Systems Biology Approaches to Somatic Cell Reprogramming Reveal New Insights Into the Order of Events, Transcriptional and Epigenetic Control of the Process

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m

(Dr. rer. nat.)

Im Fach Biophysik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I

der Humboldt-Universität zu Berlin

von

Dipl. Biophys. Till Philipp Scharp

Präsidentin/Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jan-Hendrik Olbertz

Dekanin/Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:

Prof. Stefan Hecht PhD

Gutachter/innen:

1. Prof. Dr. Dr. hc Edda Klipp
2. Prof. Dr. Ralf Mrowka
3. Prof. Dr. Nils Blüthgen

Tag der mündlichen Prüfung: 08.10.2014

It is about the real value of a real education, which has almost nothing to do with knowledge, and everything to do with simple awareness; awareness of what is so real and essential, so hidden in plain sight all around us, all the time, that we have to keep reminding ourselves over and over. This is water. This is water

David Foster Wallace, 2005

Erklärung zur Beteiligung verschiedener Autoren an präsentierten Forschungsergebnissen

Hiermit erkläre ich, dass ich diese Dissertation mit bestem Wissen eigenständig und nur mit Hilfe der hier aufgeführten Hilfsmittel verfasst habe. Zur Beseitigung von Missverständnissen sollten jedoch die verschiedenen Beteiligungen anderer Autoren an Forschungsergebnissen erläutert werden, die in dieser Arbeit vorgestellt werden. Da es üblich ist, über die Zeit der Promotion gemeinsam mit anderen Forschern an verschiedenen Projekten zu arbeiten, enthält die abschließende Dissertation, die die Forschungsergebnisse im Ganzen widergibt, immer auch Material, das vielleicht an anderer Stelle schon einmal gemeinsam publiziertem Material ähnelt. Die folgende Auflistung enthält die verschiedenen erwähnten Forschungsarbeiten und die entsprechenden Beteiligungen:

- Abschnitt 4.4: Der Abschnitt bezieht sich teilweise auf die Publikation: Bock, Scharp, Talnikar, and Klipp (2013)

  Autorenbeteiligung:

  Matthias Bock und ich waren hauptverantwortlich für die Planung und Umsetzung der im Google Summer of Code 2012 (GSoC2012) entwickelten Ideen, sowie für die Erstellung des Textes und der Abbildung für die oben erwähnte Publikation.

  Chaitanya Talnikar war hauptsächlich verantwortlich für die Umsetzung unserer Ideen in die Programmierung des Tools.

  Die Anwendung des Tools auf die in Kapitel 4 präsentierte Optimierung wurde eigens von mir durchgeführt.

- Kapitel 5: Das Kapitel bezieht sich auf die Publikation: Flöttmann, Scharp, and Klipp (2012)

  Autorenbeteiligung: Max Flöttmann und ich waren hauptverantwortlich für die Planung und Umsetzung des kompletten Projektes, der Programmierung der Skripte für Berechnung und Darstellung.

  Die Abbildungen stammen hauptsächlich von Max Flöttmann, während der Text von beiden Autoren in ungefähr gleichem Maße verfasst wurde.

Abstract

Somatic Cell Reprogramming has emerged as a powerful technique for the generation of induced pluripotent stem cells (iPSCs) from terminally differentiated cells in recent years. Although holding great promises for future clinical development, especially in patient specific stem cell therapy, the barriers on the way to a human application are manifold ranging from low technical efficiencies to undesirable integration of oncogenes into the genome. It is thus indispensable to further our understanding of the underlying processes involved in this technique.

With the advent of new data acquisition technologies and an ever-growing complexity of biological knowledge, the Systems Biology approach has seen an evolution of its applicability to the elaborate questions and problems of researchers. Using different mathematical modeling approaches the process of somatic cell reprogramming is examined to find out bottlenecks and possible enhancements of its efficiency.

How can biological networks involved in pluripotency bridge the gap between stability and plasticity through topological features? A motif analysis of a network involved in pluripotency and reprogramming revealed a striking difference in network motif abundance and stability in comparison to randomly constructed networks sharing similar network features. I hypothesize this difference to be related to sensible characteristics of iPSC networks that are involved in multi-stability lineage decisions.

What are the crucial reactions and interactions taking part in the first 96 hours of reprogramming? The optimization of a classic Boolean model gained from prior literature knowledge against early reprogramming gene expression profiles reveals new insights into the first steps of the process. In this framework, the transcription factor SP1 can be attributed a crucial task and new ideas on the wiring of critical mechanisms such as FGF2 signaling, hypoxia inducible factors and cell-cycle related functions emerge. I postulate an intermediate state in which transcriptional activity of genes playing an important role in iPSCs is strongly down-regulated.

How do epigenetic and transcriptional interactions co-operate to determine pluripotency and lineage decisions in reprogramming and differentiation and can they explain low reprogramming efficiency? A probabilistic Boolean network (PBN) of the interplay of transcription, DNA methylation and chromatin modifications, is established that aims at explaining the most important steps in the reprogramming process, tries explanations for the low reprogramming efficiencies and hints at possible enhancement strategies. Again, the aforementioned transcriptionally inactive intermediate state accumulates during reprogramming simulations.

Zusammenfassung

Die Reprogrammierung somatischer Zellen hat sich kürlich als leistungsfähige Technik für die Herstellung von induzierten pluripotenten Stammzellen (iPS Zellen) aus terminal differenzierten Zellen bewährt. Trotz der großen Hoffnung, die sie speziell im Bezug auf patientenspezifische Stammzelltherapie darstellt, gibt es viele Hindernisse auf dem Weg zur Anwendung in der Humanmedizin, die sich von niedrigen Effizienzen bei der technischen Umsetzung bis hin zur unerwünschten Integration von Onkogenen in das menschliche Genom erstrecken. Aus diesem Grund ist es unabdingbar, unser Verständnis der zugrundeliegenden Prozesse und Mechanismen zu vertiefen.

Durch neue Datengewinnungsmethoden und stetig wachsende biologische Komplexität hat sich der Denkansatz der Systembiologie in den letzten Jahrzehnten stark etabliert und erfährt eine fortwährende Entwicklung seiner Anwendbarkeit auf komplexe biologische und biochemische Zusammenhänge. Verschiedene mathematische Modellierungsmethoden werden auf den Reprogrammierungsprozess angewendet um Engpässe und mögliche Effizienz-Optimierungen zu erforschen.

Wie können Pluripotenz-Netzwerke durch topologische Merkmale die Lücke zwischen Stabilität und Plastizität schließen? Eine Motiv-Analyse eines Pluripotenz- und Reprogrammierungs-assoziierten Netzwerkes deutet auf einen signifikanten Unterschied zwischen Häufigkeiten von Netzwerkmotiven im Vergleich mit zufällig generierten Netzen hin, deren topologische Charakteristiken mit denen des Pluripotenznetzwerkes übereinstimmen. Ich vermute, dass diese Differenz auf verschiedene Stabilitätskriterien der Netzwerke hinweist.

Welches sind die entscheidenden Interaktionen, die sich in den ersten 96 Stunden der Reprogrammierung abspielen? Die Optimierung eines klassischen Booleschen Modells aus einem automatisierten Literatur-Recherche Netzwerk in Bezug auf frühe Genexpressionsdaten aus Reprogrammierungsexperimenten gewährt Einblicke in die ersten Schritte des Prozesses. Im Rahmen der Optimierung konnte dem Transkriptionsfaktor SP1 eine entscheidende Rolle zugeordnet werden und neue Ideen entstehen über die Vernetzung wichtiger Mechanismen, wie z.B. den FGF2-Signalweg, Hypoxie- oder Zell-Zyklus-Faktoren. Ich postuliere einen intermediären Zustand, in dem die transkriptionelle Aktivität einiger Schlüsselgene aus iPS Zellen herunterreguliert ist.

Wie arbeiten epigenetische und transkriptionelle Kontrollprozesse zusammen, um Pluripotenz- und Zelllinien-Entscheidungen in Reprogrammierung und Differenzierung zu treffen? Es wird ein probabilistisches Boole'sches Modell erstellt, das dieses Zusammenspiel verdeutlicht. Dabei wird versucht,

Erklärungen für die geringen Reprogrammierungseffizienzen zu finden und Optimierungen für zukünftige Experimente vorzuschlagen. Außerdem finde ich den intermediären transkriptionell inaktiven Zustand wieder, der schon vorher postuliert wurde.

Contents

1 Introduction

1.1 Embryonic Stem Cells, Induced Pluripotent Stem Cells and Aim of the Work

Since their first derivation from mouse embryos (Kaufman et al., 1983), embryonic stem cell (ESC) research has emerged to be one of the most important and most promising current fields of study in the scientific community. Discussions mainly center on stem cell therapy, ethical controversies and in recent years the term *reprogramming*. In 2012, the Nobel prize in the field of physiology and medicine was jointly granted to Sir John B. Gurdon and Shinya Yamanaka for *the discovery that mature cells can be reprogrammed to become pluripotent*, a recent discovery by Takahashi and Yamanaka (2006).

In the following, I will outline the specific characteristics of stem cells, why their use and abuse is so vividly discussed and what advantages could arise from successful stem cell therapy. Moreover, I will explain the concept of somatic cell reprogramming which makes use of genetic methods to modify differentiated cells into induced pluripotent stem cells (iPSCs) which are similar to ESCs.

1.1.1 Embryonic Stem Cells, Use and Abuse: Biological Progress vs. Ethics

Human embryonic stem cells (hESCs) are those early developmental cells that constitute the inner cell mass (ICM), also called embryoblast, in an early-stage embryo (see Figure 1.1). ESCs have a few key qualities that make them especially attractive for research. The most interesting feature consists in the fact that they can develop into every one of the three germ layers, i.e. the endoderm, ectoderm and mesoderm, and thus into all other cell lineages of the human body. This characteristic is called pluripotency as opposed to the unipotency of terminally differentiated cells. Moreover,

ESCs also have the quality to self-renew indefinitely which attributes them immortality. Taken together, these 2 key qualities make ESCs extremely interesting for research in medical as well as in biological sciences.

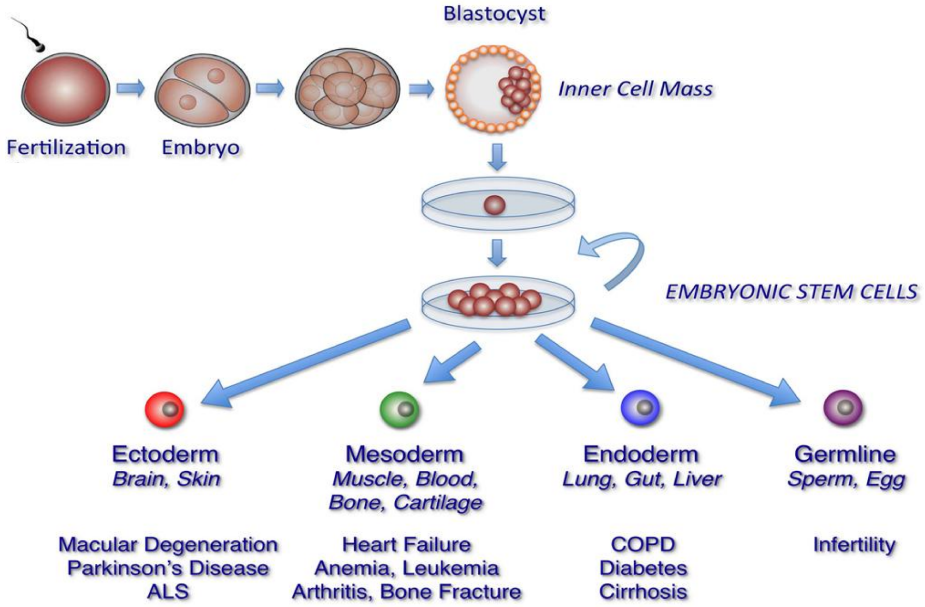

Figure 1.1: Origin and Potency of ESCs (Figure taken from Yabut and Bernstein (2011))

Embryonic Stem Cells are derived from the inner cell mass (ICM) of the blastocyst stage of a fertilized egg. When cultured and expanded in ESC medium, the pluripotency of ESCs allows them to differentiate into all 3 germ layers upon differentiation signals, i.e. ectoderm, mesoderm, endoderm and also into germline cells. At the bottom, the potential diseases are shown that could be tackled with a successful stem cell or iPSC therapy whose concept is outlined in Figure 1.2

However, there is a controversial discussion when it comes to the acquisition of hESCs, which requires the extraction of the ICM of a human embryo leading to its *death*. The discussion mainly revolves around the definition of life. At what point does a pile of cells evolve into a living creature? Is the potential of giving an organism in the future enough to talk about life? And is it wrong to artificially grow stem cells in vitro to possibly be able to help scientific progress? The debate is especially difficult to lead because of the great potential that stem cells hold for the therapy of a wealth of diseases. For a review on the ethical discussion in stem cell research please consult Lo and Parham (2009).

In fact, upon its discovery, stem cell therapy promised big breakthroughs in the cure of degenerative diseases, i.e. diseases which lead to the deterioration of the affected tissue (e.g. neurodegenerative diseases such as Parkin-

son (Lindvall and Kokaia, 2006), osteo-degenerative diseases, diabetes, etc.) (Singec et al., 2007). By today, some of the promises have held up to their potential and there is measurable progress in therapeutics with expectations from experts still being elevated for future treatments.

Another upside to stem cells, which makes them especially attractive in the field of biology and biochemistry, is their potential use as well examined, easy to handle model systems to study processes and mechanisms as well as diseases inside a cell (Jakel et al., 2004).

In summary, one can say that stem cells bear a great potential for multiple usage but there are downsides to the matter when it comes to the ethical question.

1.1.2 Somatic Cell Reprogramming as a Means to Circumvent Ethical Controversy

As mentioned above, the revolutionary discovery by Takahashi and Yamanaka (2006), that reprogramming of differentiated cells into ESC-like iPSCs upon transduction of pluripotency genes via viral vectors, shed a new light on the ethical discussion as well.

The 4 transduced genes were the transcription factors OCT3/4, SOX2, KLF4 and c-MYC. Whereas the former 2 were known to be involved in the core transcriptional regulatory circuitry of hESCs (Boyer et al., 2005), the latter 2 were rather associated with up-regulation in tumors. Approximately 16 days after infection with the viral vectors, colony formation of iPSCs that morphologically and genetically resemble ESCs could be observed.

Degenerative diseases are often due to mutations. As shown in Figure 1.2, the ideal workflow for curing such a disease using iPSCs consists in extracting any terminally differentiated cells from a patient, e.g. skin cells (Hanna et al., 2007). These cells are then treated with the four factor combination mentioned above leading to reprogrammed patient-specific iPSCs. In these latter, the genetic defect responsible for the disease can then be corrected in vitro. Afterwards, the healthy iPSCs could be re-differentiated into the cell lineage of the affected tissue and re-transplanted. This strategy offers large advantages compared to the conventional stem cell therapy where unspecific ESCs from an existing cell line would be used. In first place, since the transplanted cells originate from a patient's graft, immune rejection, which still is the biggest problem in organ or tissue transplantation, could be completely prevented. Moreover, as the iPSCs are patient specific, disease modeling and drug screening could be carried out more individually and not only for model ESCs which may not be able to reflect the vast majority of patient's genetic defects (Passier et al., 2008; hong Xu and Zhong, 2013).

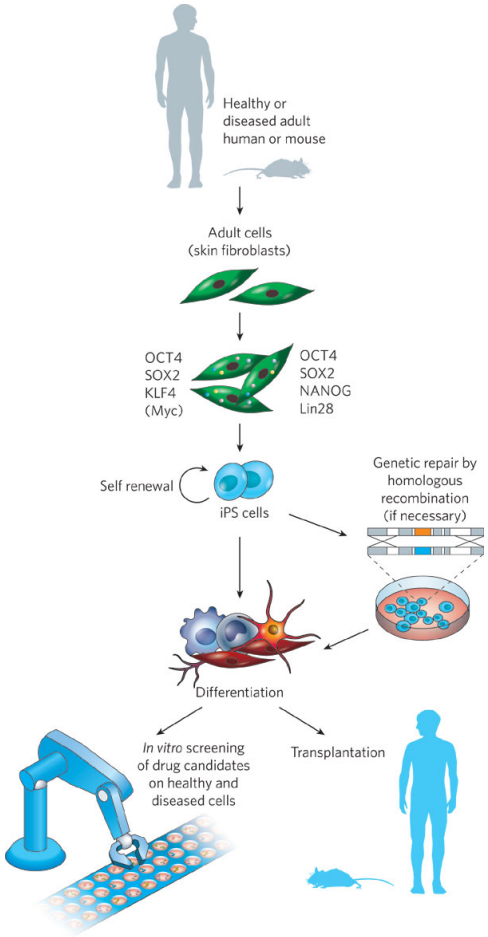

Figure 1.2: Possible iPSC therapy taken from Passier et al. (2008)

Adult differentiated cells (e.g. fibroblasts) are taken from the diseased mouse or human, cultured in a dish and transfected with the reprogramming cocktail. The resulting iPSCs are then genetically manipulated to become healthy again, differentiated back to cells of the tissue in question and re-transplanted into the patient, regenerating the unhealthy tissue

After having outlined the sequence of events of a possible iPSC treatment, one can now understand why it avoids the ethical discussion. At no point, there is a need to fertilize an egg or kill an embryo or any organisms that could be considered *life*.

1.1.3 Roadblocks on the Way to the Clinic

At first, the potential therapy described above sounds seductive and easy to put into practice. However, it should be considered that out of the four transcription factors necessary for reprogramming, three are proto-oncogenes,

namely Oct3/4, Klf4 and c-Myc (Hochedlinger et al., 2005; Yancopoulos et al., 1985; Wei et al., 2006) and with the retroviral transduction method, they will integrate into the genome. This exact method could never be used for clinical application, as upon transduction these exogenous genes are in control of a different promoter than the endogenous analogs. Hence, their expression will be uncontrollable making the transduced cells prone to a potential tumor formation. Moreover, the reprogramming efficiency, i.e. the number of cells that really form iPSC colonies is very low (way below 1%) in the process (Hanna et al., 2009). Thus, to avoid viral integration of oncogenes and improve the efficiency, new techniques have since been developed. They include transfection with plasmids (Okita et al., 2008), usage of recombinant proteins that can penetrate the plasma membrane (Zhou et al., 2009), addition of small molecules such as the histone deacetylase 1 (HDAC1) inhibitor valproic acid (VPA) (Huangfu et al., 2008) or even the very recent knockdown of Mbd3, a core member of the Mbd3/NuRD (nucleosome remodelling and deacetylation) repressor complex (Rais et al., 2013). The latter two dramatically increase the efficiency, the knockdown by Rais et al. even to 100%. The latest research breakthrough consists in the stimulus-triggered acquisition of pluripotency (STAP), a process that is claimed to be able to produce iPSCs only by applying stress such as toxins, low pH or physical pressure onto differentiated cells (Obokata et al., 2014).

Although these novel techniques are in development, the lack of understanding of the reprogramming process will make it difficult for it to be clinically applicable in the near future. In my opinion, it is thus inevitable to further study the processes by means of the Systems Biology approach which will be explained in more detail in the following

1.1.4 Understanding Mechanisms: The Systems Biology Approach

In the last century, our knowledge in the field of biology and medicine has increased by an unimaginable amount. With the advent of new discoveries, high-throughput technologies and an overall augmentation of scientific research, the field is growing more and more complex and the amount of data is expanding exponentially. It is thus a very crucial task of researchers to reconcile the vast amounts of experimental data with the underlying theories of biological systems in order to be able to draw sensible conclusions from experiments on the system level. This is where the Systems Biology approach comes into play.

Systems Biology is a way to adress complex interactions in biological systems within a more holistic instead of the traditional reductionist context. Understanding a biological system consists of understanding the topology,

the structure of the system, its dynamical behavior, how it is controlled and the relationship of its design and its function in the bigger picture. The ambitious goal of this approach is the modeling of these processes and the prediction of the system's behavior upon different stimulations or modifications. These biological systems typically are metabolic, signaling or gene regulatory networks (GRNs) (Kitano, 2002).

One of the central tools of the Systems biologist is an abstract representation of the system in question, the so-called *model*. The model of the biological system can consist of any set of compounds, e.g. genes, RNAs, proteins or small molecules inside the cell or outside of of it. These species can be represented by a set of variables describing their amount. The nature of the variables depends on the chosen modeling framework. The topology of the network, i.e. the ensemble of all species and interactions between them can be derived via exhaustive literature research, utilization of databases or design of experiments to identify interactions such as ChiP-on-chip or gene expression profiling (Chuang et al., 2010).

Having completed the network topology, it is possible to proceed to model building and dynamical analysis. This is one of the most difficult steps in the process because many questions and levels of analysis have to be considered here. It is necessary to determine the modeling framework that one should use ranging from binary Boolean modeling in different ways over different discrete and approximating continuous modeling approaches until ordinary and even partial differential equation (ODE or PDE) modeling. Moreover, this is also the point where the scope of the model has to be defined, i.e. the question that the modeling approach should answer. Is it built to gain a more detailed understanding of the processes involved or should it make predictions about possible modifications in order to enhance processes, cure diseases, estimate drug concentrations? Are we interested in the system's steady state or the exact dynamics how the system reaches these states? These questions are amongst others also determined by the availability and amount of experimental data and knowledge that can be included into the model. The main features and characteristicy of the Systems Biology approach described thus far are reviewed in Kitano (2002) and Chuang et al. (2010).

1.1.5 Scope and Aim of This Work

In order to gain a deeper understanding of somatic cell reprogramming, different modeling frameworks, optimization techniques and network characteristics will find their application in this work. I am thereby going to approach questions concerning the structure of pluripotency related networks, the most important players involved in the early stages of reprogramming

and the wiring and interplay of transcriptional and epigenetic mechanisms.

I am especially interested in the relationship between the structure and the stability of a pluripotency network involved in multi-stability processes such as lineage decisions. In the first part, I will therefore focus on network motif discovery in a gene regulatory network (GRN) of iPSCs and the relationship between network motif abundance and stability.

Furthermore, when focusing more on the process of reprogramming, the question arises which species and underlying mechanisms play the most important part and which are the first to be differentially regulated. This is why, the second part will consist of a reduction of the model used for the network motif discovery in order to integrate experimental early reprogramming data and train a Boolean model of pluripotency to it. This model will then be simulated with an in-browser Boolean simulator partly revealing the dynamics of the first necessary steps in the iPSC generation.

In order to extend our understanding from purely transcriptional interactions to the involvement and interplay of epigenetic modifications, I will add more levels of regulation in a purely theoretical probabilistic Boolean model (PBN). The analysis of this model will help to identify possible explanations for a few roadblocks of reprogramming and to find enhancement strategies.

The basics of these different modeling frameworks and mathematical theories will be outlined in Section 1.3 and Chapter 2. However, since the adaptations of the techniques for the problem sets in question will be specific and abstract, every chapter will have its own introductory part and mathematical and biological explanations will also partly be placed alongside their application in the results Chapters.

In the following Section, I will shortly outline the regulatory mechanisms of the cell that will be studied in this work. These processes are numerous and each of them is very complex. Therefore, only the basics will be treated here with a focus of the involvement of the processes in reprogramming and differentiation. A more detailed description will be effected in the respective Chapter and will then be more focused on the specific problem that is treated.

1.2 Biological Background: The Different Layers of Regulation

1.2.1 Gene Regulatory Networks and the Core Transcriptional Network of Pluripotency

The genome constitutes the template for the majority of cellular compounds such as mRNAs, the vast diversity of proteins and rRNAs in every known living organism. The central dogma of molecular biology (Watson, 1965; Crick, 1970) identifying the flow of information from genes to proteins has evolved over the decades to unravel the complex mechanisms of transcription, RNA processing and translation. One specific class of proteins are the transcription factors, proteins that are able to bind to fragments of DNA, e.g. in the promoter region of a gene, and to help (as activators) or prevent (as inhibitors) recruitment of the RNA polymerase resulting respectively in the activation or inhibition of transcription of the gene.

A gene regulatory network (GRN) is a set of genes that controls a specific set of cellular mechanisms via mutual up- or down-regulation through the transcription factors that they encode. Beside the fast acting signaling pathways that are based on protein-protein interactions (PPIs) and that will be introduced in Section 1.2.2, the GRN functional regulatory units are necessary to adequately respond to changes of external or internal conditions in order to survive or optimize protein levels at the long-term (Levine and Davidson, 2005).

It should be stated at this point that according to convention, genes and proteins in human and mouse will be coded in different manner in this work, as described in Table 1.1. However, it is clear, that when talking about a species in a mathematical model describing a biological process, it could often be either the gene or the corresponding gene product that is mentioned. Therefore, the use of italic or plain notation is often just a matter of interpretation of the sentence and the context and should not be read with absoluteness.

Table 1.1: Notation of Mouse and Human Genes and Proteins

Entity	Examplary Notation
Mouse Protein	Sox2
Mouse Gene	*Sox2*
Human Protein	SOX2
Human Gene	*SOX2*

A recently discovered example for a small GRN is the core network of pluripo-

tency master regulators that is tightly inter-connected and acts downstream on a wealth of target genes (Boyer et al., 2005). It consists of the master regulators of transcription OCT4 (transcription factor encoded by the gene *POU5F1*), SOX2 and NANOG that mutually activate each other's transcription thereby sustaining their expression once the module is active. The activity of this pluripotency module has been shown to be at the basis for the self-renewal and pluripotency characteristics of ESCs (Nichols et al., 1998; Masui et al., 2007; Mitsui et al., 2003). As shown in Figure 1.3, OCT4 and SOX2 can form a heterodimer activating their own and NANOG transcription. While OCT4 occupies around 600 downstream genes, NANOG and SOX can both bind to more than a 1000 genes. Interestingly, many of the target genes of the 3 transcription factors are shared, i.e. at least 2 out of the 3 master regulators co-occupy a wealth of target genes, thereby adding another level of regulation on to it (Boyer et al., 2005).

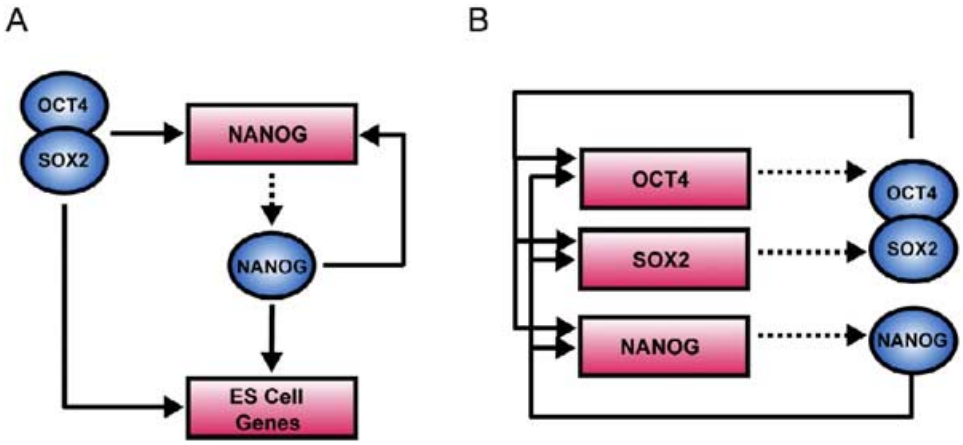

Figure 1.3: Pluripotency Core Regulatory Circuitry (Figure taken from Boyer et al. (2005))

Red rectangles represent the genes, blue circles represent the encoded proteins. **A** OCT4 and SOX2 together activate NANOG. Both parts act downstream on many target genes, thus creating a feed-forward loop. **B** The auto-regulatory core network of pluripotency. Through the mutual activations, it sustains its own expression once it is activated

One of the important cellular processes in which the master regulators of pluripotency are involved is cell lineage decision. In mouse ESCs (mESCs), artificial repression of *Pou5f1* induces trophectoderm differentiation which is regulated by a complex of Oct4 and Cdx2 which represses *Pou5f1* as well as *Cdx2* expression (Niwa et al., 2005a) and similar behavior was found in hESCs as well (Hay et al., 2004). This leads to a bi-stable system in which the decision for pluripotency or differentiation depends on the master regulators of the lineages in question. A similar mechanism between NANOG and GATA-6 is responsible for the primitive endoderm lineage decision (Mitsui et al., 2003; Niwa, 2007a; Chickarmane and Peterson, 2008) and in the

mesenchymal transcription network (MacArthur et al., 2008). This concept of bi- or multi-stability will play a role in Chapter 3 where it will be exploited to attribute stability constraints to a pluripotency network as well as in Chapter 5 where this molecular switching is a crucial mechanism for the modeling of lineage decisions that we will combine with epigenetic features in a multi-level model.

Apart from lineage decisions, the 3 master regulators of pluripotency are also involved in many other cellular processes such as cell cycle, epigenetics and signaling pathways. The latter two will be treated in the following Subsections.

1.2.2 The Role of Signaling Pathways in Human Pluripotent Stem Cells and Reprogramming

In order to survive, cells have to be able to quickly accommodate to changes in the environment. These changes can concern the availability of nutritional molecules, mating pheromones, temperature or salt concentration in unicellular organisms such as bacteria or yeast or much more complex mechanisms conveyed via hormones and other signaling molecules in higher order organisms. In order to be able to transduce the external signals into cells, signaling pathways have evolved. A membrane-bound signal receptor which can bind the signaling molecule or sense temperature or electro-physiological changes transfers the signal to cytoplasmic proteins by conformational changes and subsequent altering of the internal protein. Depending on the pathway in question the signal is passed via different other proteins from the cytoplasm into the nucleus where a transcriptional program will be activated (Berg et al., 2002). These signaling pathways also play a crucial role in the maintenance of pluripotency and self-renewal and it is well known that they are important in the processes of reprogramming and differentiation (Dalton, 2013).

I will outline the basic cross-talks of a few signaling pathways and their involvement in pluripotency related mechanisms. In fact, these mechanisms will play a role in the analyses carried out in Chapter 4.

To date, it is FGF2 signaling via the mitogen activated protein kinase / extracellular signal-regulated kinase (MAPK/ERK) pathway, Activin A, Nodal and TGFβ signaling via the SMAD2,3 branch of the TGFβ pathway, insulin/IGF signaling via phosphoinositide 3-Kinase (PI3K) and WNT signaling - the latter will not be treated in this work - that shape our knowledge of signaling in hPSCs (Dalton, 2013).

In order to sustain self-renewal, ERK has to be kept at a low level range because it quickly induces differentiation (Na et al., 2010; Dalton, 2013) at

higher levels. Contrary to earlier belief, it appears that FGF2 can maintain low levels of ERK at high or at low concentrations. This is achieved via a cross-talk mechanism between FGF2, PI3K/AKT and ERK as shown and explained in Figure 1.4. While at low FGF2 levels, ERK levels are kept at low levels as well via the signaling cascade, at higher levels, FGF2 also more strongly activates the PI3K pathway which lies downstream of IRS1. PI3K then activates AKT which in turn acts negatively on ERK thereby regulating its levels in a range favorable for self-renewal.

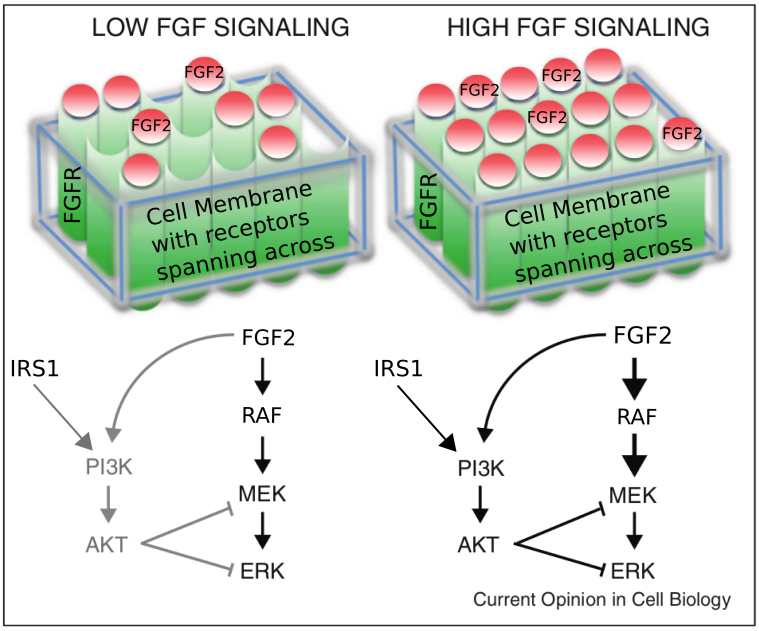

Figure 1.4: Effects of Low and High FGF Signaling

FGF2 regulates PI3K and MAPK/ERK in hESCs. In the upper part of the figure, the membrane of the cell is represented schematically with green FGF receptors (FGFR) spanning across and small red FGF2 molecules being able to bind to the receptors. Left: At low concentrations, FGF2 slightly activates MAPK/ERK signaling but keeps ERK signaling underneath a certain threshold above which it would induce differentiation. Right: At high FGF2 concentrations, another pathway is also activated: the PI3K/AKT pathway that lies downstream of IRS1. This pathway inhibits ERK activity thereby potentially regulating ERK within a range that is compatible with self-renewal (Figure taken from Dalton (2013) and extended by the IRS1 interaction which will be further explained in Chapter 4)

Another pathway that has long been known to play an important role in pluripotent cells and reprogramming is the TGFβ pathway (James et al., 2005). This pathway mainly consists of two branches, the SMAD1/5/8 (also called BMP branch) and the SMAD2/3 branch (also called TGFβ branch), activation of the former leading to differentiation and the latter sustaining pluripotency and self-renewal (Greber et al., 2008). However, it was also found that reprogramming to iPSCs requires a mesenchymal-epithelial transition (MET) (Samavarchi-Tehrani et al., 2010) which is inhibited by the

TGFβ branch and favored by the BMP branch of the pathway that favors an epithelial-mesenchymal transition (EMT) (Li et al., 2010). Taken together, these results seem contradictory at first, because the TGFβ branch is related to pluripotency and self-renewal but blocks the MET necessary for the reprogramming and thus the transition to pluripotency from differentiated cells.

This shows that there are controversial results when it comes to signaling pathways and their relationship to pluripotency and reprogramming. In fact, the interpretation and analysis of the mechanisms of action of signaling pathways is highly sensitive to the employed culture conditions, the isolated observation of the pathway instead of its integration in the cellular context and the level of activation of the pathway as mentioned before with the ERK regulation via FGF2 and PI3K. Moreover, it should be noted that especially the signaling pathways have different roles in mESCs and hESCs (Schnerch et al., 2010) and their interpretation should therefore be treated with the highest care. A more complex intertwining and possible cross-talking between the different pathways will be given alongside the discussion of the results in Section 4.3.

1.2.3 Epigenetics: The Extended Dogma of Cell Biology

In contrast to transcriptional and signaling pathways regulatory mechanisms, epigenetics constitute a more restrictive and thus higher level of regulation. The term *epigenetics* was first used by C.H. Waddington in the concept of the *epigenetic landscape* (Waddington, 1942, 1953). He therein developed a framework to describe the loss of potency of differentiating cells comparing them to bowls rolling down a hill in a ragged landscape (see Figure 1.5). The point of highest elevation of this landscape represents the pluripotent state while the points of lower elevations represent the less potent state passed by the cell in order to arrive in a terminally differentiated state at the bottom of the hill.

New definitions of the term followed much later by Holliday (1990) and Riggs et al. (1996) relating epigenetics to gene activity and heritability independent of the DNA sequence. The most recent consensus definition states it as the *stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence* (Berger et al., 2009). These changes, called *epigenetic modifications*, can affect multiple structures of the chromosome in various ways. In the chromosomes, DNA is associated with histone proteins to form the highly condensed chromatin responsible for DNA packaging, mitosis and the control of gene expression. Epigenetic modifications either affect the DNA molecule (without changing the sequence) or the aforementioned histone proteins. While for the DNA, the main modification consists

in cytosine methylation, for histone proteins, many of the modifications are known, e.g. methylation, acetylation, phosphorylation or ubiquitination (for review see Bártová et al. (2008)). There are specific enzymes that can transfer the modifying chemical groups onto the molecules such as Histone Methyl Transferases (HMTs) (Wood and Shilatifard, 2004) and others that can remove them again such as Histone Deacetylases (HDACs) (reviewed in Sengupta and Seto (2004)). Since epigenetic modifications are tightly related to transcriptional control, a complex mutual regulation of these processes takes place inside the cell.

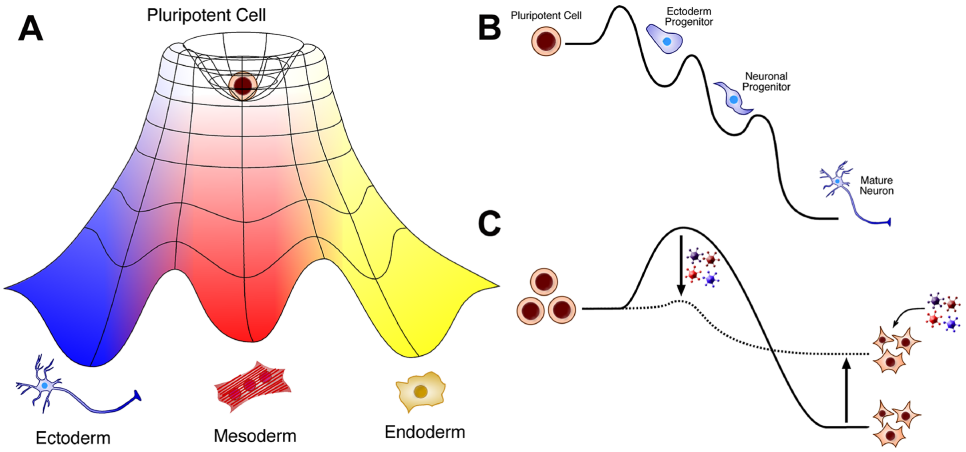

Figure 1.5: The epigenetic landscape and its implications for direct reprogramming (Figure and caption taken from Rodolfa (2008))

A A Waddington-inspired schematic of the epigenetic landscape. Culture conditions will promote the self-renewal of a pluripotent cell, maintaining it in a shallow well at the top of a cellular potential hill. When allowed to differentiate, this cell will "roll" down the hill into one of many terminally-differentiated fates at lower potential. **B** A closer look at the path a pluripotent cell might take as it differentiates into a neuron, passing through a number of intermediate progenitor states of varying stability on the way. The line in (B) represents a slice through the surface shown in (A). **C** The process of direct reprogramming, like chemical catalyst, implicates a restructuring of the epigenetic landscape. Introduction of the transcription factor cocktail destabilizes the fibroblast identity while stabilizing the transition state. Because the retroviruses are shut down in the iPS cells, however, the potential of the pluripotent state remains unchanged

In this work, I will mainly focus on DNA (de-)methylation, histone (de-)methylations, histone (de-)acetylations and their interplay. These modifications can effectively alter the transcriptional activity of the genes that are affected by the modifications. In which direction the modifications alters the transcription, either in an activating or an inhibiting sense, strongly depends on the modification, the affected residue of the modified molecule and the context. The detailed mechanisms relating epigenetics, especially DNA methylation and differential chromatin structures upon histone modifications will be outlined alongside the creation of our multi-level model in Chapter 5.

There have been quite a few modeling efforts on the subject of pluripotency and somatic cell reprogramming. Therefore, in the following Section, the mathematical background, i.e. the state of the art of these models will be outlined.

1.3 Mathematical Background: Pluripotency and Somatic Cell Reprogramming in Models

It is very complicated to fully understand the effects and consequences of the complex interplay of the above mentioned regulatory processes. This is where mathematical models can help to resolve the order of events and put together the cellular behavior and its link to the underlying molecular mechanisms. Since processes involved in reprogramming could in theory span everything that happens inside of a cell, an enormously complex system, it is necessary to reduce the amount of information in order to determine and evaluate the basic features underlying the behavior of the network. Different publications have adressed the modeling of certain parts of more or less complicated regulatory networks with valuable success (Kalmar et al., 2009; MacArthur et al., 2008; Chickarmane and Peterson, 2008; Saez-Rodriguez et al., 2007).

The thus far described regulatory mechanisms only work perfectly together when they are executed in an orchestrated fine-tuned manner. Previous publications have described quite a few networks regulating pluripotency in stem cells and during reprogramming. They partially explain the bistability of the system decisions taken in development and the influence of expression noise (Chickarmane et al., 2006; Chickarmane and Peterson, 2008; MacArthur et al., 2008; Kalmar et al., 2009). What all of these models have in common is the application of ordinary differential equations in order to reveal the dynamical features of a small subnetwork of the whole regulatory machinery inside the cell. Larger networks have recently been modeled using the dynamic Bayesian networks approach which suggested improved reprogramming factor combinations (Chang et al., 2011).

Furthermore, more coarse grained models have been developed in order to describe transitions between cell states and self-organization in the cell (Halley et al., 2009; Qu and Ortoleva, 2008). However, those models are very abstract, based on a conceptual approach and don't describe single genes such as the pluripotency master regulators and their synergy. In an earlier work seeking to analyze chromatin remodeling, Dodd et al. (2007) showed the necessary existence of a positive feedback during heterochromatin formation.

When looking at the ensemble of experimental and theoretical efforts de-

scribed thus far, a strong evidence emerges that reprogramming requires a stochastic component that drives the process in a directed manner. Very valuable insights into the relationship of proliferation rates, reprogramming times and efficiency have been gained by modeling reprogramming as a stochastic process of one simple state transition with a corresponding probability distribution (Hanna et al., 2009). In the first modeling approach including epigenetic features and transcriptional regulation into a mathematical model of reprogramming, Artyomov et al. (2010) designed the ensemble of developmental states as a binary decision tree where nodes represent cell states with the pluripotent state at the base of the tree from which the other originate. This study even offered an explanation for the low reprogramming efficiency. The probabilistic Boolean model that I will present in Chapter 5 has a little similarity to this latter model. However, it uses a different modeling approach, includes more detailed mechanisms and goes in a different direction.

2 Materials and Methods

2.1 Biological Methods

2.1.1 Microarray Gene Expression Profiling of Early Reprogramming

As mentioned earlier, recent years have seen an explosion of high-throughput technologies generating a vast amount of experimental data. One of these techniques is the gene expression profiling using DNA microarray chips with the ability to simultaneously measure thousands of genes at the same time. In this case, the measured quantity is the gene activity or expression, i.e. the relative quantity of mRNA in an assay. The ensemble of gene expression values, the so-called expression profile, contains a high amount of information on the instantaneous state of the cell. In fact, a cell has the theoretical abilty to produce all mRNAs and proteins that are encoded by the genes on its DNA. However, in a given state and point in time, it only transcribes a small fraction of all these genes, its transcriptional profile, which is dictated by the transcriptional, signaling and epigenetic mechanisms that were briefly mentioned in Subsections 1.2.1, 1.2.2 and 1.2.3.

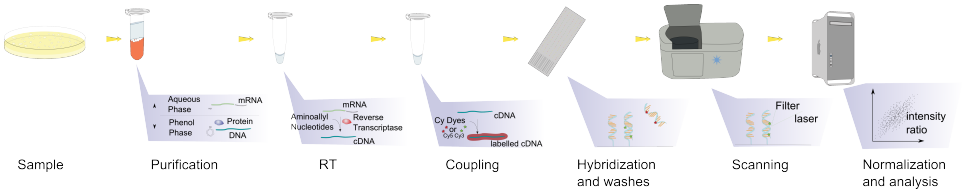

Figure 2.1: Microarray Gene Expression Profiling Experiment

DNA microarray chips consist of an ensemble of small fragments of DNA arranged on the surface of a chip in so-called DNA spots via covalent binding to a solid phase. These DNA spots are used to hybridize complementary DNA or RNA (cDNA or cRNA), which is DNA or RNA that was obtained

via reverse transcriptase catalyzed copying of a certain mRNA and thus is complementary to that latter. The thus gained cDNA is then labeled with fluorophores and its hybridization on the DNA chip via complementary nucleic acid sequence binding is detected via fluorescence measurement. Analyzing the strength of the fluorescence signals of every spot in comparison to a background signal yields a specific amount of bound target onto the spot via a complex normalization procedure that will be described in an example in the following Subsection. A schematic representation of a microarray experiment is represented in Figure 2.1.

2.1.2 Raw Data Analysis of Early Reprogramming Microarray Gene Expression Profiling Data

The following analysis has been carried out by Dr. Guifré Ruiz-Acero in his Ph.D. thesis (Ruiz Acero, 2012)

Human fibroblasts were transduced with viral vectors containing different transcription factors, namely OCT4, SOX2, KLF4 and c-MYC in 6 different assays: The first 4 assays contain fibroblasts transduced with only one of the 4 transcription factors, while assay 5 and 6 are combinations of 3 transcription factors (assay called 3TF comprising the genes *OCT4, SOX2, KLF4*) or all 4 transcription factors (assay called *4TF*). Another assay contained a viral vector carrying the *GFP* gene as a control (assay called *GFP*) beside the control without any transduction (assay called *FIB*). After 4 days (96 hours), cRNA was hybridized to DNA-microarrays as explained in Subsection 2.1.1 in 3 biological replicates of all the assays. The replicates were then averaged and differential expression analysis was carried out for every assay. In this analysis, the *GFP* measurement is considered as background measurement.

The complete expression data analysis uses the software BeadStudio 3.0 by Illumina (http://www.illumina.com/). The raw microarray data are background-subtracted and normalized with the *rank invariant* algorithm. In order to filter for differentially expressed genes, normalized data are subsequently compared to the *GFP* control. The computed fold changes are then selected for differentially expressed genes considering genes as up-regulated if the signal intensity ratio $Factor_{assay}/Factor_{GFP} > 1.5$ and down-regulated if the signal intensity ratio $Factor_{assay}/Factor_{GFP} < 0.67$. Only highly confident signals with a detection p-value < 0.01 are considered as differentially expressed.

In Chapter 4, these thus found differentially expressed genes will be used to filter a big interaction network. The raw data of the genes will moreover be normalized using the rescaling method that is described in Subsection

2.2.5 in the following and be used to train a Boolean network of the filtered interaction network.

2.2 Software

2.2.1 Cytoscape

Cytoscape is an open source software tool for the visualization and data enrichment of complex networks. There is a wealth of plugins available and it has very practical features for different fields of research, e.g. bioinformatics, systems biology or genomics (Shannon et al., 2003). These features include:

- Loading molecular and genetic interaction data sets in various standard formats such as *.sif, .gml, .sbml*, excel or delimited text files
- Enrich networks with experimental data or annotations
- Network analysis and export of the results
- Create visual mappings based on data or network analysis results
- Layout the network with a wealth of layout algorithms

to only name the few most used features in this work.

Cytoscape 2.8.2 was used to create Figures 3.1, 4.1, 4.2, 4.3, 4.4, 4.5, 5.3, 5.4, 5.5.

2.2.2 Genomatix Pathway System (GePS)

The Genomatix Pathway System (GePS) from the Genomatix company (http://www.genomatix.de) is an in-browser software tool for the storage and generation of biochemical pathways. It uses information from public and privat databases to create interaction networks based on complex automated literature mining algorithms and subsequent expert curation (Frisch et al., 2009). The iPSC core network version 2 that is stored in the software is used in Chapters 3 and 4 for the discovery of specifically enriched network motifs and the training of a reduced Boolean model version of it to early reprogramming microarray gene expression profiling data.

2.2.3 Python

I used the programming language python (http://www.python.org/) together with many of its packages in the following parts of the thesis:

I generally used *networkx*, *pyparsing* and *re* packages for parsing and manipulating network files and writing conversion scripts for different graph formats in Chapters 3 and 4.

I used the *os*, *subprocess* and *random* packages for the automatization of random Boolean network generation and motif detection in Chapter 3.

2.2.4 mFinder

The tool mFinder (Kashtan et al., 2004) is a software released by the Weizmann Institute for the detection and statistical analysis of network motifs containing from 2 up to 6 nodes. The algorithm for the full enumeration of these subgraphs is described in Milo et al. (2002). The tool is used for the motif discovery and partly the statistical analysis in Chapter 3.

2.2.5 R

The statistical software environment *R* (http://www.r-project.org/) was used intensely throughout the thesis to carry out statistical tests and generate, analyse and optimize Boolean models using various software packages as will be outlined in the following.

Statistics

The statistical tests in Chapter 3 were carried out using the *base* package of *R* for the Shapiro-Wilk, the Bartlett, the Welch, the Wilcoxon-Mann-Whitney tests as well as the Student's t-test and the *car* package for the Levene test.

BoolNet

BoolNet is an R package for the generation, analysis and visualization of Boolean networks (Müssel et al., 2010). In this work, the package is mostly used in chapters 3 and 5. In the former, which treats of the network motifs discovery in random and pluripotency related networks, it is employed for the generation and subsequent analysis of random Boolean networks (RBNs). It is especially useful to look for attractors and their basin sizes and to filter the RBNs for these criteria in Chapter 3. In Chapter 5, which summarizes our publication Flöttmann, Scharp, and Klipp (2012) it is used to conduct a Markov simulation of a probabilistic Boolean network (PBN). Especially the transition matrix $\mathbf{A}$ of the Markov process (which will be introduced further below in 2.3.4) is generated using the corresponding method of the BoolNet

package. The visualization of the 3-dimensional time course in Chapter 5 was carried out using the *persp3d* function of the R *rgl* package to create a 3-dimensional landscape plotting the probability over time for each of the $r = 2^n$ states as can be seen in Figure 5.7 in Chapter 5.

CellNetOptimizer and a few Extensions

The *CellNOptR* package for *R* is a software tool that integrates topological data in prior knowledge networks (PKNs) and experimental perturbation data in order to optimize a Boolean model of the PKN (Terfve et al., 2012). Originally designed for protein signaling networks, it is used in this work for the training of a transcriptional interaction network involved in pluripotency to microarray gene expression profiling data of early reprogramming under various conditions. Since it is extensively used in this work in Chapter 4 and since its mechanism of action is complex and a few changes have been applied to it, it will be thoroughly explained in the following.

As a first step, the PKN and the experimental data set need to be converted into data structures that are accepted by the software. For the PKN, this is the Simple Interaction File (SIF) format, while for the experimental data the Minimum Information for Data Analysis in Systems Biology (MIDAS) is the format of choice. The SIF format is a pure interaction format, in which nodes and the edges between them are specified in order to build a graph. An example of such a file can be deduced from Table 4.1, in which the entries of the first column are source nodes, the second entry is the type of interaction in which a "1" signifies activation while a "-1" designates inhibition and the third entry is the target node. The data are given to the software in the MIDAS format which specifies the measurement condition, time point and species that is measured in a tabular form (For more detail on the MIDAS format pleas consult Saez-Rodriguez et al. (2008)).

Subsequently, in order to run the training of a Boolean model against continuous microarray data, the latter needs to be normalized somehow between 0 and 1. The software offers 3 ways of carrying out this normalization:

1. In the *CTRL* mode a fold change at the same time and same experimental condition with respect to a control assay is taken. Such a control was not measured in our data set.

2. In the *Time* mode, a fold change with respect to the time point 0 is taken and normalized via a complicated procedure described in Saez-Rodriguez et al. (2009) and in the Appendix in Section A.2. However, the normalization procedure in this case always transforms the initial condition to 0 and then computes positive values for an increase of the species' concentration and negative values for a decrease. Naturally,

the outcome of a Boolean model can never be negative, which is why species that decrease in the data set can only be reflected by species in the model that are 0 at the beginning and stay 0. This is an undesirable bias in the optimization because in fact, inhibition of an expressed species can never be described.

3. The *raw* mode applies the same procedure as the *time* approach but it is the raw values that are transformed via the described method and not the fold changes. The advantage is, that the method does not transform the data at time point 0 to 0 constantly. However, there is another problem with this approach. In fact, the normalization procedures includes a transformation via a Hill function in the following way:

$$\frac{x^{HillCoeff}}{EC50Data^{HillfCoeff} + x^{HillCoeff}} \tag{2.1}$$

where x is the respective data point, *HillCoeff* is the Hill coefficient used for the normalization and *EC50Data* is the normalization parameter corresponding to half-maximal saturation in Hill kinetics. However, the parameters for the normalization, especially the *EC50Data*, are taken as equal across all species, although the species have very different concentration values. Using the same parameter for all values is meaningless and error-prone.

For these reasons mentioned, I carried out the normalization procedure in Section 4.2 manually. I chose rescaling as a means to normalize data continuously between 0 and 1 applying the following equation:

$$\frac{S^i - S^i_{min}}{S^i_{max} - S^i_{min}} \tag{2.2}$$

where S^i is the concentration of species i, S^i_{min} is the minimum concentration of species S^i across all conditions and time points and S^i_{max} is the maximum. Equation 2.2 is carried out for every species at every condition and time to transform every data readout into a value normalized between 0 and 1. I deliberately refrain from discretization in this context. In fact, discretization diminishes the content of information of the data by assuming the existence of binary states that might not exist in reality. Indeed, there will be intermediate states: some genes might already be expressed but could be down- or up-regulated upon different stimuli. Rescaling the data will account for this qualitative behavior while discretization would insist on the existence of binary states that are either *ON* or *OFF*. Therefore, the rescaling method is the method of choice to reflect biological reality in a better way.

Following the data processing, the PKN SIF file is treated in several steps to build an ensemble of logic models which make up the state space for optimization. The network processing consists of a compression and an expansion step: The compression step eliminates species that are not measured or perturbed to reduce the model complexity. It is still necessary, however, to keep the complete PKN in mind to map back the optimized model at the end in order to identify which nodes and which edges are necessary or very likely to be present to fit the data. The expansion step transforms the topological network into a set of Boolean models: In fact, for each node, all possible logic gates for the inputs (or Boolean functions) are created. As an example, if a node C has two possible input nodes A and B, the expansion will create the 4 possible gates, that is A activates C, B activates C, A and B are necessary to activate C, A or B are necessary to activate C. The latter two of these Boolean functions and their possible molecular basis are represented in the introduction in Figure 2.4. This is just one molecular example to describe Boolean OR and AND gates which can in fact account for a wealth of possible underlying mechanisms that involve 2 or more input species that affect a target species.

Every one of the thus created possible logic gates for every node gets assigned a bit in a bit string (or bit vector) that fully describes the model. A "1" at the specific position of the bit vector means the corresponding logic gate was present in the optimization, a "0" means it was absent. The goal of the optimization process is to search through the vast state space of all of these possible model structures (or possible bit strings) and find out the ones that fit the data best with the possible outlook to draw conclusions on the molecular mechanisms that is imposed by the trained model. It is the bit vector of fixed length described above that is optimized during the process. The optimization function (or score of the optimization or objective function) is shown in the following equation:

$$\frac{1}{n}\sum_{t,l,k}(M_{t,l,k} - D_{t,l,k})^2 + \alpha\frac{1}{s}\sum_{edges} e_{edges} + \beta n_{NA} \tag{2.3}$$

where n is the number of data points, i.e. number of species multiplied by the number of measured time points times the number of conditions for that time point, $M_{t,l,k}$ and $D_{t,l,k}$ respectively the values of the model output and the the measured data point for readout (species) l and condition k at time t, α is the size factor that penalizes the edge term which is composed of the sum over all edges in the optimized model normalized by the total number of hyperedges s and finally β is the NA factor that penalizes the number of undetermined values n_{NA} returned by the model. The model for which to compute the value $M_{t,l,k}$ is obtained by translating the bit string of logic gates into a Boolean model structure. It is important to notice that the

model output $M_{t_{end},l,k}$ corresponds to the value of species l at condition k after the model has reached its steady state at the second time point t_{end} ($t_{end} = 96h$ in the microarray data set which is represented in the appendix in Section A.1).

There are a few parameters that are common to genetic algorithms that will be explained in the following. The *Population Size* for each generation of the evolutionary algorithm is the number of models randomly generated per generation and their corresponding value of the fitness (or optimization) function. The *Probability of Mutation* describes with what probability a solution taken from the last generation is slightly changed to generate a new result while the *Elitism* parameter determines how many of the best solutions of the last generation are taken into the next generation unchanged. Moreover, there are 3 parameters that are able to stop the optimization. The *Maximum Number of Stall Generations* is the number of consecutive generations in which the the best score and the model (the bit string to optimize) can stay the same before the algorithm stops. The *Maximum Time* and *Maximum Number of Generations* are respectively as the names state the time (in seconds) and number of total generations that the algorithm runs before stopping. In all the optimizations that will be run and discussed in Chapter 4, the *Maximum Number of Stall Generations* is always the factor stopping the optimization. The optimizations were designed in a way that after 300 generations of unchanged best results, the actual best solution for the problem is hypothesized to have been found. The *Selective Pressure* measure is a slightly more complex means to rank the solutions and its exact description can be found in Bäck and Hoffmeister (1991) and Whitley (1989).

In fact, in every optimization step, a certain number of models (the population size which I chose to be 100 as declared in Table 4.2) are generated by the genetic algorithm and every one of them is simulated until a steady state is reached. This steady state value is taken as the model output value. If a model doesn't reach a steady state, e.g. in the oscillatory case, a "NA" is generated.

During the optimization, the software tool keeps track of the parsed models and saves the ones that have a score within the tolerance interval of the best model that can be defined by the user. This ensemble of models is then used to compute the weight of edges, i.e. the probability of an edge to be present in the model which is nothing else than the relative frequency of models inside the tolerance where the edge in question is present.

For better understanding, I will quickly outline the derivation of this probability. I have introduced earlier the bit vectors to fully describe a model. Every bit in these bit vectors corresponds to a Boolean logic gate, a "1" or "0" corresponding to the gate being present or absent respectively in the model. These bit vectors should not be mistaken for the bit vectors describ-

ing the state of a given model! If we consider a set A of k models with bit vectors of length n and A_{ij} the j-th bit of model A_i with $i \in \{1..k\}$ and $j \in \{1..n\}$, we can define the relative frequency of occurrence N_{ij}^{rel} of the logic gate corresponding to this j-th bit in all models of the set A as:

$$N_{ij}^{rel} = \frac{1}{k} \sum_{i=1}^{k} A_{ij} \tag{2.4}$$

In fact, every time the gate is present, the value of A_{ij} will be 1 while it will be 0 when it is not present. Therefore, the sum exactly counts the number of occurrences in the set A and division by the total number of models k yields its relative frequency of occurrence or probability. It is very important to notice that due to combinatorial constraints inside the bit vector as well as to the model size (α) and model output (β_{NA}) constraining expressions in the optimization function, the expectation value for N_{ij} is far below 0.5. This relative frequency of occurrence will be used in Section 4.3 for the filtering of optimized edges and the graphical representation of the results.

In order to test the similarity of models inside the tolerance interval of one optimization and across different optimizations, I computed a similarity score for each optimization based on the bit strings describing the models that have been mentioned before. The relative difference D between two models of one optimization can just be described by taking the sum of the square of the difference of the two bit vectors and dividing it by their length L:

$$D = \frac{\sum_i (A_i - B_i)^2}{L} \tag{2.5}$$

where A_i and B_i are the i-th entries of the bit vectors **A** and **B** respectively. Since A_i and B_i can only take the values 0 or 1, the difference measure defined above is nothing else than the relative amount of bits that differ in the two vectors. Therefore, if we want to know how similar the two bit vectors are, we just have to subtract the measure from 1 yielding the similarity $S = 1 - D$.

CNO is used exclusively in Chapter 4 and will be further explained alongside its application in this chapter.

2.3 Mathematical Methods

2.3.1 Statistical Hypothesis Testing

At some point in the course of a classic Systems Biology approach, it will be necessary to assess how experimental data are distributed, how well a model works, how well it fits the data and to which degree certain hypotheses are valid or not. For many biophysical problem sets such as model inference, fitting or comparison of different data sets with certain distributions, statistical hypothesis testing is the analysis tool of choice. In general, the method checks whether results are statistically significant or in other words likely not to have occurred randomly by chance alone by testing whether a certain hypothesis is accepted or rejected at a certain level of significance. The mechanism of action of a significance test will be briefly outlined in the following.

In first place, it is important to preliminarily define the *null* or *alternative* hypothesis which can differ from one specific test to the other. Next, it is crucial to define the nature of the distribution, i.e. its characteristics concerning shape and variance, in order to determine the relevant statistical test that has to be carried out. The chosen statistical test than calculates a test statistic and a p-value the latter corresponding to the probability of observing at least the obtained test statistic under the assumption that the null hypothesis is true. This p-value is then used to accept or reject the null hypothesis at a certain significance level α that is usually chosen to be 0.05 or 0.01. I will now quickly outline a workflow of hypothesis tests used for the comparison of 2 samples that will be useful in Chapter 3.

In order to test for similarities in two samples, I designed a hypothesis testing decision tree that suits my needs, that takes into account the assumptions on the distributions in question and that helps to determine the statistical tests that need to be used for the problem (see Figure 2.2). The decision tree presented here only takes into account 2-sample tests which are used to compare two different samples. The first test that needs to be carried out for the two samples is the Shapiro-Wilk test for normality. It tests the null hypothesis that a sample drawn from a normal distribution against the alternative that it is not. This first decision determines the next step in the tree. In the normal case, a Bartlett test (Bartlett, 1937) is carried out to test whether the two samples have the same variance while in the non-normal case we have to use the Levene test for the same purpose (Olkin, 1960). Two normally distributed samples with the same variance can be compared in the famous Student's t-test (Student, 1908) while the Welch adaptation to the t-test is used for samples with same variance (Welch, 1947). In the non-normal case, when the two samples happen to have the

samee variance, the Wilcoxon-Mann-Whitney (or Mann-Whitney-U) test can be applied (Wilcoxon, 1945; Mann and Whitney, 1947) while samples with different variances can be compared with the very general and not very powerful Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948). For a more detailed explanation of statistical hypothesis testing please consult Lehmann and Romano (2005).

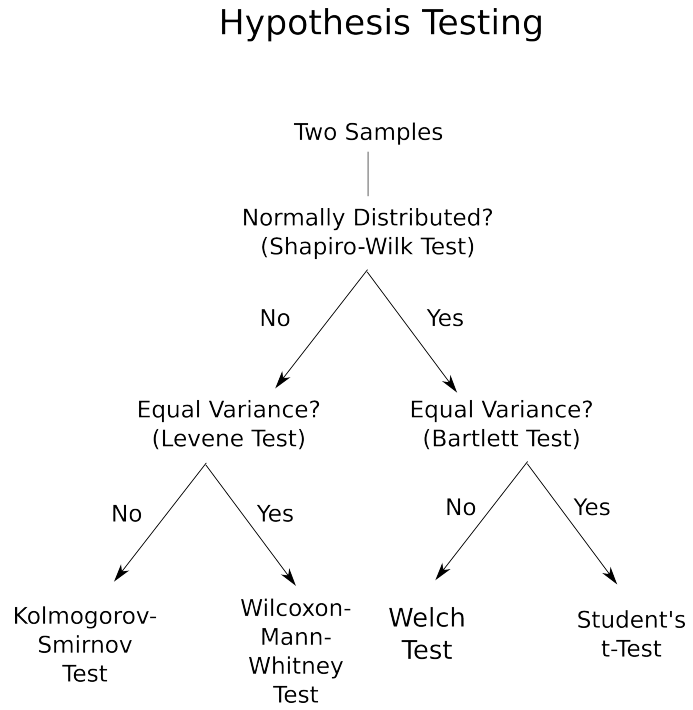

Figure 2.2: Decision Tree for Hypothesis Testing

This binary tree implements the decision making process described in the text in Subsection 2.3.1 and will be employed in Chapter 3. Starting from two sample distributions, it describes how they are progressively tested for normality, equal variances and finally equal means or medians in order to assess whether the two samples are likely to have been drawn out of the same distribution

A slightly more detailed depiction of the tests to carry out and their progressive work flow will be described alongside their application. In this work, statistical hypothesis testing will play a crucial role in Chapter 3, where different distributions of network motifs - that will be presented in the following Subsection 2.3.2) - and their relative frequencies will be tested for equality or difference in the multi-step decision process that I presented above (see Figure 2.2).

2.3.2 Network Motifs: Detection and Dynamic Behavior

In order to approach big interaction networks governing pluripotency, it is important to understand static and dynamic features of the network topology and their apparent relationship. Regulatory networks, such as transcriptional interaction networks, need to be tightly regulated in order to be able to regulate target genes upon external changes. The high complexity and dynamics of interactions are only slowly being uncovered. It has thus recently been found that small regulatory patterns involving a certain defined number of nodes, e.g. 3 or 4, occur significantly more often than expected in biological networks. These small subgraphs were called network motifs (Shen-Orr et al., 2002; Milo et al., 2002) and their expectancy was calculated by searching their occurrence in randomized networks (Milo et al., 2002).

The 13 3-node network motifs that were found are shown in Figure 2.3 with their IDs as used in the *BoolNet* package of *R*. Interestingly the motif IDs uniquely describe the underlying motif by transforming a long binary integer of the motif's adjacency matrix into a decimal number. The feed-forward loop with ID 38 for example (see Figure 2.3) is described by the adjacency matrix:

$$\begin{array}{cccc} & A & B & C \\ A & 0 & 1 & 1 \\ B & 0 & 0 & 1 \\ C & 0 & 0 & 0 \end{array} \tag{2.6}$$

where every row and every column represents one of the nodes A, B or C and the entries of the adjacency matrix describe whether a directed interaction exists (1) or not (0). When concatenated, this adjacency matrix yields the binary number 011001000 or 38 in the decimal system.

There are several algorithms to find network motifs in large networks including exhaustive search, algorithms based on sampling, scalar subgraph counting amongst others that are reviewed in Ciriello and Guerra (2008). I used the method implemented in the mFinder tool that is described in Milo et al. (2002) and Kashtan et al. (2004).

While the structure of a network as well as the analysis of network motifs is static, we are more interested in the dynamical behavior of the system. Since the exact network behavior is dictated by the interplay of all species and thus by the exact topology of the network, a relationship between static and dynamic features cannot be denied. However, an exact analysis of the dynamics of a large network with many interactions is a very exhaustive process and takes a lot of time and effort. At the beginning of an interaction

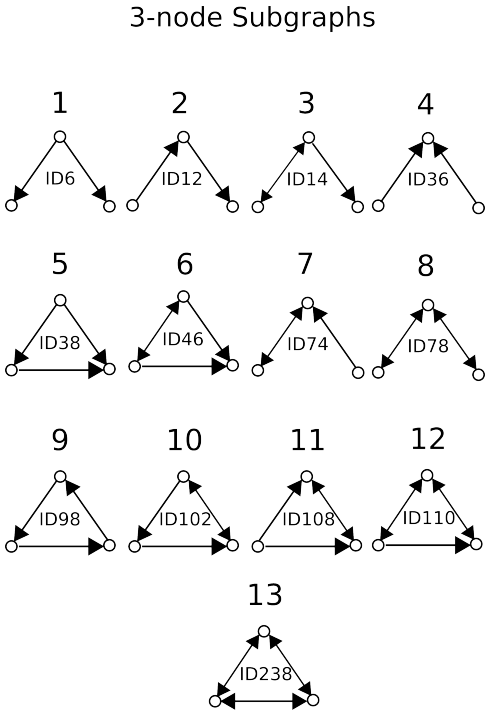

Figure 2.3: Dictionary of All 3-Node Motifs With Their IDs as Used in the R BoolNet Package (derivation of motif ID see plain text)

network study, one may only be interested in a qualitative insight into the general network dynamics. The functional relationship between abundance and dynamics of single network motifs with the behavior of the whole system is thus a compelling analysis. In order to assess this relationship, it is necessary to translate single motifs into a dynamical system. The most interesting feature of such a system is then the analysis of the stability of its steady state, i.e. its reaction to perturbations which can be stable, unstable or oscillatory depending on the type of the interactions in the motif (activating or inhibiting) and their strength.

Therefore, Prill et al. (2005) defined a metric, the structural stability score (SSS) to assess the stability of every network motif. It describes the probability that upon small perturbations the mathematical system built to describe the network motif in question relaxes monotonically to the same steady-state. The SSS is comprised between 0 and 1 with a value of 1 corresponding to non-oscillatory, stable behavior over a wide parameter range, i.e. monotonic relaxation to the steady state while the lower the value, the more the parameter range of the dynamical system for stable behavior is narrowed down. The exact derivation of the SSS can be found in Prill et al. (2005). The important thing to notice is that the SSS classifies motifs into 3 cate-

gories of stability: While highly stable motifs of class (I) with an $SSS = 1$ lack feedback loops, the second group of motifs with an $SSS \approx 0.4$ contain exactly one feedback loop and low stability motifs with an $SSS < 0.2$ consist of more complicated subgraphs with different combinations of multiple feedback loops.

In Chapter 3, I will analyze the motif frequency of 3-node subgraphs in a pluripotency related interaction network in comparison to random networks and compare their difference with respect to the SSS defined above. The SSS of the 13 3-node motifs is nicely displayed in Figure 3.3 alongside the results of that chapter. For the generation and analysis of the random networks, I will use random Boolean networks (RBNs). Boolean networks will moreover play an important role throughout the complete thesis which is why they will be introduced in the following Subsection.

2.3.3 Boolean Logic and Modeling: A Binary View on Biological Systems

Boolean algebra is a mathematical concept named after George Boole who was the first to approach logic in an algebraic way in his 1847 work *The mathematical analysis of logic*. His ideas were further developed by mathematicians such as Jevons, Whitehead, Schröder, Stone and a few others until the 1960s to the Boolean algebra that we know today. The application of Boolean logic to biological networks, first described by Stuart Kauffman in the 1960s (Kauffman, 1969), will be the major mathematical approach to modeling, simulating and optimizing biological networks in this work.

Since Kauffman (1969) introduced random Boolean networks (RBNs), or also called Boolean NK networks for biological simulations, they have been further developed and are extensively used to describe dynamic network behavior (i.e. cell cycle (Waltermann et al., 2010), signalling (Saez-Rodriguez et al., 2007) or stem cell differentiation (Flöttmann et al., 2012)). In the concept of RBNs, N is the number of nodes representing the system's variables that take binary values and K was initially a fixed number of input edges per node but can also be a set of numbers varying for each node. The random character of RBNs comes from their first introduction by Kauffman (1969) who constructed networks by randomly associating input genes to every node in the network thus creating a randomized network or an ensemble of randomized networks that could thus be simulated and their characteristics evaluated.

Although Boolean models very strongly simplify biological reality, they are still a very useful tool to qualitatively examine dynamical network behavior especially in extended networks or when approaching young research fields

where exact knowledge of model parameters and data are scarce or only qualitatively available. Moreover, they have proven useful and generated good results in the past when applied correctly especially in the field of developmental gene regulatory networks (Macía et al., 2009; Kauffman, 2004).

In Boolean networks nodes can only take the discrete values 0 and 1 sometimes also called *True* and *False*. Biological entities are thus exclusively viewed as active or inactive, phosphorylated or unphosphorylated, expressed or not expressed, carrying specific molecular features such as epigenetic marks or not carrying them and nothing in between. This is a simplification of reality, since biochemical species can be present in a continuous range of concentrations. However, signals are often transmitted via threshold crossings and all-or-nothing decisions. Thus it is often appropriate to discriminate between two states - below and above the threshold - which have different qualitative characteristics.

Transitions between these states are defined by Boolean logical operations that can account for activation, inhibition and more complicated cooperative interactions. These logical operators describe interactions between the species of a network and are defined using Boolean algebra (or logical operators) such as AND-, OR-, NOT-gates (or mathematically written as $\wedge$, $\vee$ and $\neg$). In a very simple example, if transcription factors A and B are both needed to activate a gene C, we would put this into the logical function $C(t+1) = A(t) \wedge B(t)$. A molecular example of this logical function is described in Figure 2.4. The use of Boolean networks in mathematical modeling of biological processes is also justifiable by its practicality and the insights gained by it in many publications (Fauré et al., 2006; Saez-Rodriguez et al., 2007; Waltermann et al., 2010; Flöttmann et al., 2012).

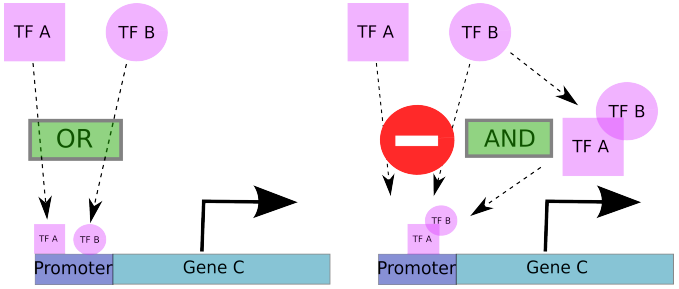

Figure 2.4: Molecular Mechanisms that can be reflected by Boolean functions

On the left hand side are shown two transcription factors A and B that can bind independently to the promoter of a target gene C to induce its expression. On the right hand side, another mechanism of DNA binding is displayed that can account for the Boolean AND gate connecting two input genes (or transcription factors) to a target gene: Sometimes transcription factors need to dimerize first, before being able to bind to the promoter of a gene or they can bind independently but the target gene can only be transcribed when both are bound.

I will now go into more detail of the mathematical background underlying

different types of Boolean networks. A Boolean network can be represented as a graph $G = (V, E)$ consisting of a set of n nodes (or vertices) $V = \{v_1, ..., v_n\}$ and k edges $E = \{e_1, ..., e_k\}$ representing interactions between these nodes. Each vertex v_i has a defined state $v_i(t) \in \{0, 1\}$ at every time point $t \geq 0$ for $i \in \{1, ..., n\}$, representing an active property of species i for $v_i(t) = 1$ or an inactive property for $v_i(t) = 0$. In a classic Boolean network (CBN), the *state vector*, or simply called the *state* of the network $\mathbf{S}(t) = (v_1(t), ..., v_n(t))$ is the vector of the node states at time t. This state vector is also sometimes described as a bit string, i.e. a string consisting of 0s and 1s instead of a vector. Since every vertex is binary and thus can only have 2 possible values 0 or 1, the total number of possible states is $r = 2^n$. This is the state definition for CBNs. At every discrete time point t, the network state is updated following a set of Boolean update functions $F = \{F_1, ..., F_n\}$. To be more precise, every function F_i defines a new value for the state of node $v_i(t)$ at time $t + 1$. Every function F_i integrates the input information on one node, i.e. how the other nodes are influencing it. Therefore, the update functions are functions of the m_i input nodes of each node with $m_i \in \{0, .., n\}$ at time t. The number of possible input functions for one node is $2^{2^{m_i}}$ and thus increases double exponentially with the number of inputs to this node. For better understanding, all possible input functions for a node with $m_i = 2$ inputs are shown in Table 2.1.

Table 2.1: Truth Table for all Possible Combinations of $m_i = 2$ Boolean Variables A and B

Every column of 4 digits on the right side from the double vertical line represents one different Boolean function. The table demonstrates the existence of $2^{2^{m_i}} = 16$ Boolean Functions for $m_i = 2$ input nodes. The rule number at the bottom is the decimal translation of the binary value of the function read from top to bottom

Input		Output															
A	B																
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Rule		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

A time course simulation of a Boolean network is the succession of states from a starting state as the discrete time progresses. At every time point t, the update function of the state $\mathbf{S}(t)$ will create a state $\mathbf{S}(t + 1)$ which can either be different from $\mathbf{S}(t)$ or the same. In the latter case, the state $\mathbf{S}(t) = \mathbf{S}(t + 1)$ is called a *point attractor* and the set of states that after a certain time lead to this attractor is called its *basin of attraction*. In Boolean networks, these point attractors are one of two classes of steady states (or attractors). The other class is the one of the *cyclic attractors*. A cyclic

attractor is a set of nodes in which all nodes will be visited again after a certain time T called the *period* with the characteristic: $\mathbf{S}(t) = \mathbf{S}(t+T)$. The remaining states are either passed exactly once in a CBN simulation, the so-called *transient states*, or are *leaf states* that can never be reached during a simulation unless they are the starting state.

All the $r = 2^n$ states of a Boolean network with n nodes make up the state space of the network. This state space can be regarded as a directed graph $Q = (S, T)$, where S is the set of states $S = \{\mathbf{S_1}, ..., \mathbf{S_r}\}$ and T is the set of edges between the states, the so-called transitions. A time course simulation can then be visualized as a path through this state space graph until it reaches a node with a self-loop that cannot be left again and is thus a point attractor. A cyclic attractor can easily be discovered as a cycle in the graph visualization. Since Boolean networks often have more than one attractor, every basin of attraction in the state space is visualized by a separate subgraph in which all states lead to the steady state of this basin. The whole state space graph will consist of as many subgraphs as there are attractors. It is important to notice that such a directed state space graph with only one output edge for every node is only the representation of synchronously updated CBNs, while in an asynchronously updated CBN's state space, the nodes can have many outgoing edges and in probabilistic Boolean Networks (PBNs) the situation is even more complicated. When dealing with Boolean networks, it is very important to distinguish between the network graph and the state space graph.

Another feature that can be derived from the graph representation of the Boolean state space is the stability of point attractors in the network. In order to fully understand this concept, let's remember that every state in a Boolean network is a succession of n binary digits of 0 or 1 (or bits), thus also the steady state. The stability of a point attractor quantifies the behavior of this attractor upon perturbation, i.e. upon spontaneous changing of 1 or more of its bits. If the perturbed state over time returns to the steady state that was initially perturbed, it accounts for the stability of the attractor, if not, it accounts for instability. Since a synchronous Boolean network is deterministic and perturbations are generally made at random points of the attractor's state vector of bits, the basin size of the attractor is a direct measure of its stability. In other words, if a steady state is perturbed, the probability for the perturbed state to be part of the basin of attraction of the unperturbed attractor linearly increases with the basin size of the attractor and can just be written as its relative size in the whole state space:

$$P = \frac{S_{attr}}{2^n} \tag{2.7}$$

where S_{attr} is the basin size, i.e. the number of nodes that are part of the

subgraph of the attractor and lead to the attractor after a certain time. This feature will serve as a stability criterion in Section 3.2.

I will now introduce probabilistic Boolean networks that will be applied in Chapter 5.

2.3.4 Probabilistic Boolean Modeling

The following Subsections 2.3.4, 2.3.5, 2.3.6 and 2.3.7 are all referring to Chapter 5 and are based on our publication Flöttmann, Scharp, and Klipp (2012).

In the beginning, Probabilistic Boolean Networks (PBNs) were designed to represent the lack of knowledge as to which regulatory Boolean functions would best represent the underlying molecular mechanism. For example, if there is experimental evidence that transcription factors X and Y bind to gene Z, but it is unclear whether they will have an activating or an inhibiting effect or whether they can bind separately or only in combination, this can be expressed by using not only one Boolean function, but a set of functions to describe the interaction. This is the main assumption of PBNs: Every node v_i doesn't have one fixed update rule F_i as in the CBNs from the introduction in Subsection 2.3.3, but its state is defined by one or more functions. The function F_i from CBNs is thus replaced by a set of functions $F_i = \{f_j^i\}$ with $i \in \{1, ..., n\}$, $j \in \{1, ..., l(i)\}$, where f_j^i is a Boolean logic function and $l(i)$ the total number of functions considered for node v_i. As mentioned earlier in Subsection 2.3.3, the total number of possible Boolean functions for a node with m inputs is $T = 2^{2^m}$ which is why $l(i)$ has to be in the set $\{1, ..., T\}$. Each of the functions f_j^i gets attributed a probability $p_j^i \in [0, 1]$ with which it will be chosen at any given point in time. A PBN can be considered as an ensemble of N standard classic Boolean networks, where $N = \prod_{i=1}^{n} l(i)$.

As in CBNs, a PBN also has the $r = 2^n$ states of the CBN that can be reached. We define a probability distribution vector $\mathbf{D^t} = (D_1^t, ..., D_r^t)$ over these $r = 2^n$ states at each time point t. Each element D_h^t (with $h \in \{1...r\}$) of the vector $\mathbf{D^t}$ corresponds to the probability of the network to be in state $\mathbf{S_h}$ with $h \in \{1, ..., r\}$ at time t.

A PBN is a time-homogeneous discrete Markov process, i.e. a process with discrete time steps, a finite state space and a transition matrix that is constant over time. The latter is defined as a $(r \times r)$ matrix $\mathbf{A}$, that contains all the probabilities to transition from state $\mathbf{S_g}$ to state $\mathbf{S_h}$ for all $g, h \in \{1, ..., r\}$. In the case where there is no network allowing the transition $\mathbf{S_g} \rightarrow \mathbf{S_h}$, the matrix entry $A_{gh} = 0$, otherwise A_{gh} is the sum of the probabilities of all the networks allowing this transition.

Another possibility to simulate a Markov process is to run a stochastic simulation over a great number of runs and then averaging the results. In this way, in every run one CBN out of the set of CBNs that constitutes the PBN is chosen at random and simulated over time. However, the advantage of calculating the transition matrix compared to stochastic simulations with a large number of runs consists in the fact, that if we choose a $(1 \times r)$ vector $\mathbf{D^0}$ with a start probability for each state we can make use of the algebraic features of geometric progressions to recursively simulate the system from t to $t+1$ (Equation 2.8) or as well directly deduce the value at a time point of interest $t+1$ (Equation 2.9):

$$\mathbf{D^{t+1}} = \mathbf{D^t} \cdot \mathbf{A} \tag{2.8}$$

$$\mathbf{D^{t+1}} = \mathbf{D^0} \cdot \mathbf{A^{t+1}} \tag{2.9}$$

The vector $\mathbf{D^t} = (D_1^t, ..., D_r^t)$ now comprises the probabilities of all $r = 2^n$ states at time t, i.e. the probability of the network to be in this state. When carrying out these calculations over a certain number of time steps, we thus get the evolution of the system's probability distribution over time. In other words, we know at every time point in which state the system will be and the corresponding probability, which is to say we get a *probabilistic time course* of the PBN. The initial conditions for the simulation can be wide-ranged: from a single state to a broad probability distribution over several states. It will be clarified further below in Subsections 2.3.6 and when presenting the results of Chapter 5, how a distribution of states might be closer to biological reality than a single state. The Markov simulation can also be used to determine the stationary states or attractors of the system as well as states that have a high transient probability. The visualization of a probabilistic time course and especially the ordering of the different states is much more complex than the one of CBNs which is why it will be explained in more detail in the following

2.3.5 Sorting Boolean States by Closeness to Template States: A Similarity Matching Algorithm

The system in Chapter 5 consists of 16384 states whose random plotting would be confusing and prevent any meaningful conclusion. Therefore, we defined a measure for the distance between states to sort them in a relevant order on an axis. I defined a measure that groups together states that are similar and put more different states in a distance to each other. The resulting 3-dimensional figure represents to the so-called epigenetic landscape (although in our model there is also a transcriptional influence) of the cell (Figure 5.7).

As mentioned above, we plot the 3-dimensional landscape with the 16384 states along the x-axis, time steps of the simulation on the y-axis and the state probabilities along the z-axis in Chapter 5. Since we are dealing with a Boolean network, the entries of every binary state vector $\mathbf{S_g}$ are obviously either 0 or 1. We apply a specifically designed two-step similarity matching algorithm to every state in the network to carry out the sorting.

To test our model in Chapter 5 and reproduce certain experimentally known dynamical behaviours, reprogramming and differentiation experiments are simulated. In these latter, we have to define at which positions in the landscape states that are similar to differentiated states or to the pluripotent state are lying. Therefore, we first have to define the state vector for fully reprogrammed cells and for the two differentiated cell lineages. This is done by deducing which variables have to be active or inactive for the system to clearly be in the state in question. We call these template states $\mathbf{S_1}$, $\mathbf{S_2}$ and $\mathbf{S_3}$. To refine the matching algorithm, we also define 3 weight vectors $\mathbf{W_1}$, $\mathbf{W_2}$ and $\mathbf{W_3}$ that attribute a certain weight to every binary variable in the template state vectors that depends on the importance of the variables for the integrity of the state. The state and weight vectors are defined in table 5.2 in Chapter 5.

We will now outline how the sorting algorithm works in detail to characterize every state in the state space. We call *matching vector of two state vectors* $\mathbf{S_g}$ and $\mathbf{S_h}$ the vector $\mathbf{M_{gh}}$ which contains a 1 for every binary variable that is identical in both vectors $\mathbf{S_g}$ and $\mathbf{S_h}$ and 0 for the ones that are different:

$$\mathbf{M_{gh}} = (\delta_{(S_{g1}S_{h1})}, ..., \delta_{(S_{gn}S_{hn})}) \tag{2.10}$$

where S_{gi} is the i-th element of vector $\mathbf{S_g}$ (with i still being element of $\{1, ..., n\}$ as defined in the beginning) and δ_{xy} is the Kronecker delta with x and $y \in \{0, 1\}$ defined by

$$\delta_{xy} = \begin{cases} 1 & \text{for } x = y \\ 0 & \text{for } x \neq y \end{cases} \tag{2.11}$$

After this, we now define the *specific similarity* σ_{gw} of a state $\mathbf{S_g}$ to one of the three template states $\mathbf{S_w}$ ($w \in \{1, 2, 3\}$), as the scalar product of the weight vector $\mathbf{W_w}$ with the matching vector of the two states $\mathbf{M_{gw}}$

$$\sigma_{gw} = \mathbf{M_{gw}} \cdot \mathbf{W_w} \tag{2.12}$$

When calculating the specific similarity to each of the 3 template states for every state, we obtain 3 sets of specific similarities σ_{i1}, σ_{i2} and σ_{i3}.

Unfortunately, the values in these 3 sets are strongly overlapping up until now, i.e. they contain approximately the same numbers. This is due to the fact, that the specific similarity is only a scalar measure of the distance between two vectors which can and will be the same for many distances of different states to the template states. Therefore, the specfic similarities are now weighted and summed up as shown in equation 2.13 to visually separate them in the landscape representation, yielding the sorting score Σ_i^{123} for every state:

$$\begin{aligned}\Sigma_i^{123} =& a * \sigma_{i1} * (\sigma_{2,max} - \sigma_{i2}) * (\sigma_{3,max} - \sigma_{i3}) \\ &+ b * \sigma_{i2} * (\sigma_{1,max} - \sigma_{i1}) * (\sigma_{3,max} - \sigma_{i3}) \\ &+ c * \sigma_{i3} * (\sigma_{1,max} - \sigma_{i1}) * (\sigma_{2,max} - \sigma_{i2})\end{aligned} \tag{2.13}$$

with a, b and c being parameters - to assign distinct orders of magnitude to the 3 sets of states - which can be tuned. Instead of just summing up the 3 weighted specific similarities, we also introduce correction terms for each of them. We define a maximal specific similarity $\sigma_{j,max}$ which is attributed to the template states themselves. Since the matching vector for this specific similarity will be a vector just filled with 1's, this maximal specific similarity simply corresponds to the sum of the elements of the template state vector in question. The correction term increases the efficiency of the algorithm regarding the separation of distinct and clustering of similar states on the x-axis. Plotting the simulation landscape, color-coding and re-distributing the states according to this sorting makes it possible to discriminate between states and makes tendencies in reprogramming and differentiation experiments visible. Apart from this clustering, when moving between states in the epigenetic landscape, they will get more similar when moving towards a template state and more different when moving away from it which will be important to notice when looking at the reprogramming and differentiation simulation in Chapter 5 (see Figure 5.7).

2.3.6 Boolean Start States and Start Distributions

As seen above, a Boolean system with n species, has 2^n states. Since our main model in Chapter 5 consists of 14 nodes, it can take on $2^{14} = 16384$ binary states. When simulating a model over time, it is necessary to define initial conditions for the time course. It should be taken into account that a cell population, even if we restrain us to one cell lineage only, can be represented not only by one state but by an ensemble of states that are similar to each other. This is due to genuine biological fluctuations, genetic and epigenetic variability and different environmental factors. However, the

specific configuration that perfectly characterizes the cell lineage in question which consists of the master regulators being expressed and epigenetic marks unset while the other modules are not expressed and their silencing epigenetic marks are set, is unique. This is the state which most of the cells of the population will be in while the other states that are similar and can also be present in the lineage have a lower probability to be attained. This latter feature can be represented by a normal distribution around the optimal state. Hence, to implement such a distribution, we create the vector of initial state probabilities $\mathbf{D^0}$ by randomly generating a normal distribution around the template state. The other states affected by the distribution are assigned probabilities depending on their similarity to the template state. This is how the distributions for the plots in Figure 5.6 in Chapter 5 were created.

2.3.7 Paths Through the Probabilistic Boolean State Space

The state space of a Boolean network with n nodes, independent of its nature as a classical, asynchronous or probabilistic Boolean network, includes 2^n states. The ensemble of all these states and the possible transitions between them can be represented as a directed graph with 2^n nodes. While in classical synchronous Boolean networks, which are deterministic, every node has exactly one incoming and one outgoing edge corresponding to the state transitions, in probabilistic Boolean networks, every state can reach every other state with a certain probability between 0 and 1 and thus there can theoretically be up to 2^n possible transitions from every state. The visual representation of such a state space, beside being very exhaustive, will also be very little instructive. Since we are mainly interested in the reprogramming process, we start the simulation from a certain set of states and only consider states that are attained with at least a certain minimum probability (see Figure 5.3 in Chapter 5).

3 Network Motif Analysis of Pluripotency Related Networks Yields a Significant Accumulation of Structurally Unstable Motifs

3.1 Significant Differences in Motif Frequencies Between Random Networks and an iPSC Network are Related to Structural Stability

As outlined in Subsection 2.3.2, there are certain network topological characteristics, the so-called network motifs, that can be related to dynamical features. In the first steps of approaching a biological interaction network that is involved in defined processes, it is interesting to be able to gain quick insights solely by analyzing its topology. Especially in the field of stem cell research, pluripotency and somatic cell reprogramming, the knowledge about relationships between structure and function is still scarce. Therefore, the first aim of this study consists in the analysis of network motifs frequencies in a regulatory network involved in pluripotency and processes in iPSCs in comparison to random networks that share general network topological features with the iPSC network. This analysis has the goal to relate topological with possible dynamical features of the networks in order to unravel certain characteristics of the iPSC network.

The latter is taken from the Genomatix Pathway System (GePS), a database that contains interaction graphs involved in cellular processes that were constructed via automated literature mining followed by expert curation. The

exact algorithms at the basis of this tool are described in Frisch et al. (2009) (see Subsection 2.2.2). The network, the tool and the company itself can be found at http://www.genomatix.de/. Using the network analysis tool of Cytoscape (see Subsection 2.2.1), the extracted network containing 125 nodes revealed to have an average in-degree of 3.90625, i.e. an average number of incoming edges to each node (see Figure 3.1). This latter measure is an important feature of interconnected regulatory networks and will later serve as a parameter in the generation of similar random networks. The exact gene list of the network as well as all interactions are only shown in the overview in Figure 3.1 in this chapter since they don't play a role for the topological analysis. However, as the network will be further used in the next Chapter 4, a reduced version of it will be shown with all genes and interactions in Table 4.1.

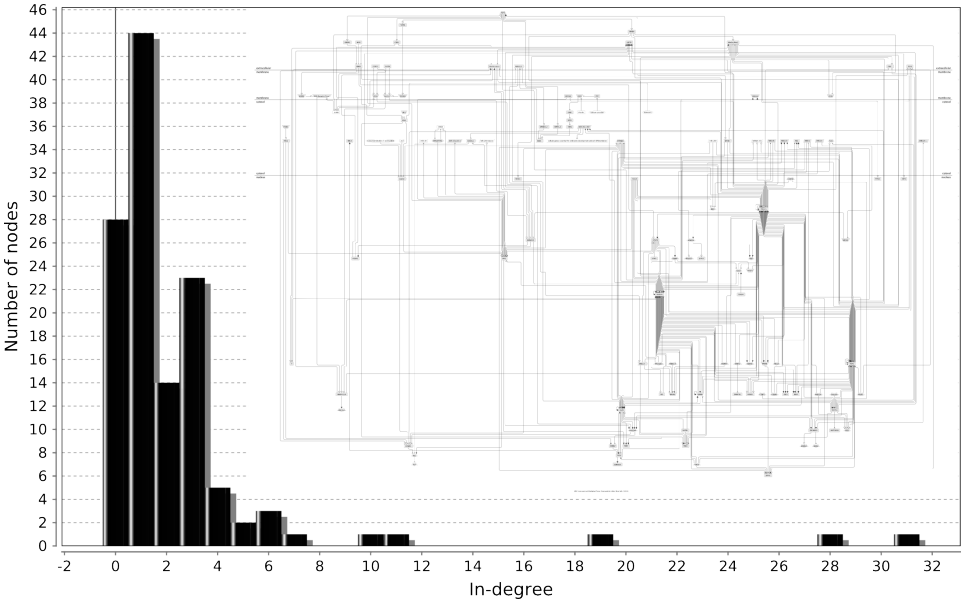

Figure 3.1: iPSC Network With its In-Degree Distribution

The big iPSC Network from the automated literature search and expert curation by Genomatix is schematically represented in the upper right. Due to its size, its species and interactions cannot be displayed. However, one can already recognize the 3 big hubs corresponding to the pluripotency master regulators OCT4, SOX2 and NANOG. The network is embedded in the graph of its in-Degree distribution with the number of nodes on the y-axis and the corresponding in-degree, i.e. number of nodes that are inputs to the node in question, on the x-axis.

The random Boolean networks (RBNs), that are described in detail in Subsection 2.3.3, were generated with the BoolNet package of R (see Subsection 2.2.5). This allows us to generate random Boolean NK networks where the parameters n, the number of nodes of the network and k, a parameter describing the topology of the network, can be fixed in advance. Although it would suffice in this Section to generate random networks and not Boolean

networks, the latter will be employed in Section 3.2 of this Chapter. Therefore, in order to keep a consistency throughout this Chapter, the random networks in this Section have also been created using BoolNet. A set of 100,000 different random networks that share important features with the iPSC network have thus been constructed using the *generateRandomNKNetworks()* method of the BoolNet package. The similarity between the random Boolean networks (RBNs) and the iPSC core network from Figure 3.1 is reflected by setting the parameters for the number of nodes $n = 125$ and for the network topology $k = 3.90625$. The latter is the parameter of the Poisson distribution of the random variable x in equation 3.1 from which the number of input nodes for every node is drawn independently to construct a homogeneous network topology.

$$F(x,k) = \frac{k^x * e^{-k}}{x!} \tag{3.1}$$

The thus created 100,000 random networks were subsequently analyzed with the mFinder tool (see Subsection 2.2.4) in order to find out the frequencies of the 13 3-node subgraphs (see introductory Section 2.3.2) in every network. For every motif, the frequency in every single random network was extracted and normalized to the number of motifs present in the network in question to yield a relative frequency of the motif in this one network. This is done for all 100,000 random networks yielding 100,000 relative frequencies which are automatically binned and the frequency of the bin across all 100,000 random networks is plotted as a histogram showing the distribution of relative frequencies of a single motif across 100,000 random networks (Represented for all 13 motifs in Figure 3.2). Moreover, the relative frequency of the motif in question in the iPSC network is plotted as well. However, this relative frequency is only one value and would not be visible in the histogram because it neither has an extension on the x-axis where the bins normally are group of values, nor in the y-dimension because it is only one value. Therefore, in order to make it visible, the relative frequency is artificially over-represented in both dimensions, i.e. assigned an artificial frequency on the y-axis and an artificial width on the x-axis, and colored in red. The differences in appearance of the artificial red bar are due to the differences in appearance of the distributions, i.e. of the bin (or breaks) width and number that is automatically generated in an optimal manner using the histogram plotting function of R. However, the bars were chosen to yield approximately half of the maximal frequency of the strongest bin and half their width. The thus created histograms for the 13 motifs are shown in Figure 3.2.

After having assessed the topological features of the network, the next step consists in relating them to possible dynamical features. As described in the introduction in 2.3.2, Prill et al. (2005) constructed a measure that relates

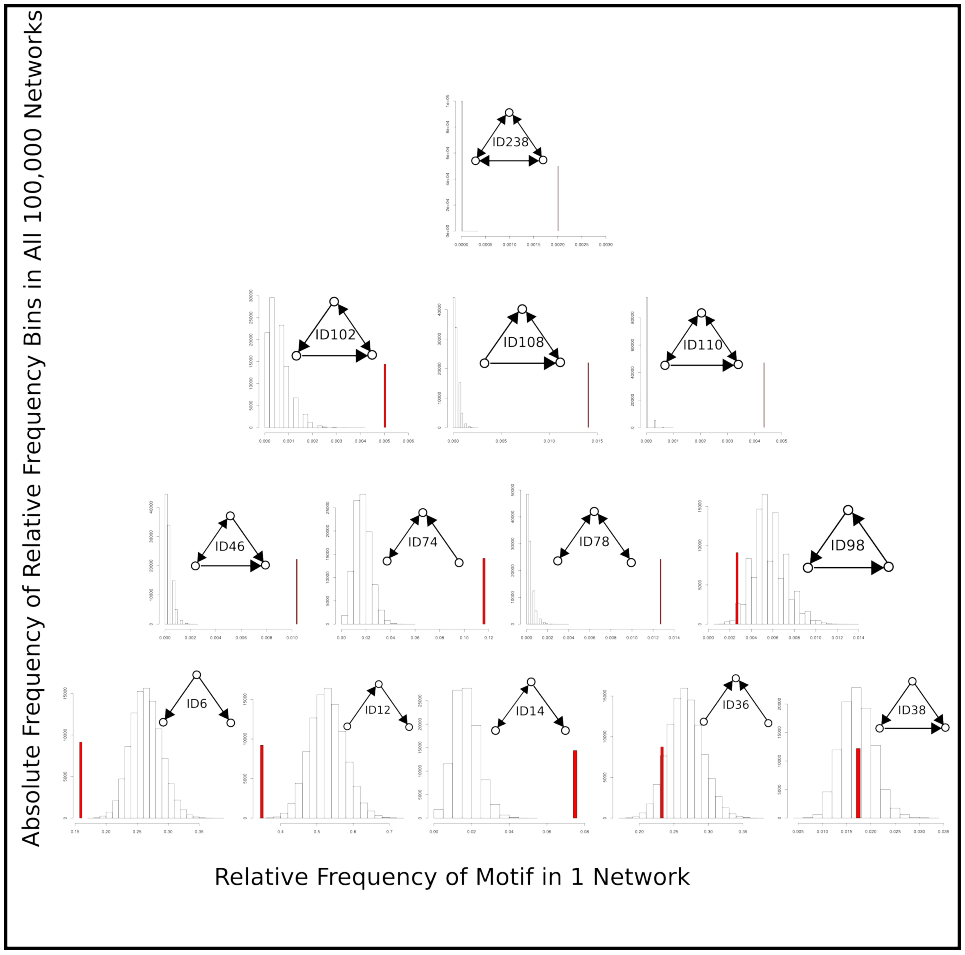

Figure 3.2: Motif Distributions of Random and iPSC networks

For all 13 motifs that are represented with their structure and their ID as well, the distribution of the relative frequency of the motif in the 100,000 125-nodes random networks is represented. The x-axis represents the bins of relative frequencies of the motif in question in every one random network while the y-axis represents the absolute counts of these relative frequency bins across all 100,000 RBNs. Together with the distributions of motifs in the random networks, the relative frequency of motifs in the iPSC core network is represented as a red bar. In order to be visible, it gets assigned an artificial count on the y-axis which corresponds approximately to the half-maximal count of the distribution. This representation clearly demonstrates that most motif frequencies in the iPSC network strongly lie outside the distributions of the RBNs. This can be seen as well in Figure 3.3.

micro-topology such as 3- or 4-node subnetworks (or motifs) with dynamical features, i.e. the characteristics of the time-dependent response to perturbations of the system. They called this measure the structural stability score (SSS). A SSS of 1 is related to structural stability, i.e. a guaranteed relaxation of the system to the perturbed steady state while decreasing values indicate progressive instability to perturbations. In order to relate the SSS to the motif distributions, I represented the mean and standard deviation

of all 13 3-node motifs across the 100,000 random networks as well as the relative frequency of the motifs in the iPSC network together with the SSS of the 13 motifs in Figure 3.3. The following interpretation is based on a graphical analysis, while a more detailed statistical analysis will be carried out in the next Section 3.2 and will support the results of the graphical analysis. It can be seen that while motifs with a SSS of 1 (Motifs with ID 6, 12, 36 and 38) are significantly under-represented in the iPSC network (as was also deduced by the values lying significantly outside the distributions in Figure 3.2), medium stable motifs with an SSS of 0.4 (Motifs with ID 14, 46, 74, 108) are significantly over-represented and motifs with a very low stability score are slightly over-represented although the small number of motifs might disallow a significant conclusion in this area.

Although it is controversial whether the ensemble of building blocks of a network, the motifs, directly influences the dynamic of the system as a whole and determines its function (Savageau, 2001; Ingram et al., 2006), it can still be hypothesized that some network motifs possess characteristics that play an important role for the function and thus are getting over- or under-represented throughout evolution thereby determining the structure. The link between structure, function and dynamics, although necessarily present in my opinion, might, however, be very complex and difficult to understand. In this work, it will be hypothesized that the frequency of a motif in a network has the ability to influence the stability of the system as a whole depending on its SSS.

As for all biological networks, a network that governs pluripotency must also have a certain stability. However, as described in the introduction in Subsection 1.2.1, it should be able to quickly leave its pluripotent steady state upon certain differentiation triggering perturbations in a multi-stable switch like behavior. This concept of multi-stability will be further outlined in the next Section 3.2. It is thus reasonable to hypothesize, that motifs that are present in a pluripotency network possibly confer less stability to it than in other networks. Interestingly, as can be seen from Figure 3.3, this seems to be the case here. However, although the hypothesis strongly correlates with our analyses, drawing a definite conclusion on the causal relationship would probably lead to far here.

Moreover, it should be mentioned that the *homogeneous* option was used to construct the RBNs while large biological interactions network have been shown to be *scale-free* (Barabasi and Albert, 1999; Barabási and Oltvai, 2004). The homogeneous condition has been mentioned further above and is shown in Equation 3.1 while for a scale-free network the degree the condition is shown in Equation 3.2 where P(k) is the fraction of nodes with k connections to other nodes in the network and γ is the scaling parameter which in the majority of cases lies in the interval $2 < \gamma < 3$.

$$P(k) \propto k^{-\gamma} \tag{3.2}$$

One of the most important features of a scale-free network is the low in- or out-degree of the majority of nodes, i.e. most nodes have only one or very few incoming and outgoing edges, while a few nodes have a very high, way above average, in- and/or out-degree, so called *hubs*. It is well-known and can moreover be seen in Figure 3.1 that some of the transcription factors, the so-called master regulators of pluripotency, regulating a wealth of downstream target genes, have a very high out-degree and thus are hubs of the network. It was shown before that biological networks are scale-free because this feature increases their robustness and thus protects them from random failure in the case of mutations or otherwise induced loss of nodes in the network (Zhu et al., 2007).

Taken together, this reveals that the difference in the motif frequencies in the 100,000 RBNs and the iPSC core network, beside possibly being an indicator for the decreased stability of the attractors in the pluripotency regulating network, might also account for the increased robustness of biological, scale-free networks in comparison to homogeneous random networks. It is at this point indispensable to understand the difference between stability and robustness which apparently, against all lexical intuition, can have opposing trends. While the stability of an attractor, as described above, is a measure for the behavior of a steady state towards small perturbations, robustness qualifies the behavior of the system as a response to structural changes of the model.

As a short summary, I have found that the iPSC network has a motif distribution that significantly differs from random homogeneous networks in such a way that network motifs conferring stability are under-represented while network motifs that are less stable and add dynamical plasticity to the system are over-represented. These features could possibly be related to the function of a network involved in pluripotency that needs to have multi-stability with at least one of the attractors having a decreased stability. This concept will be extended in the following

In the following, beside applying hypothesis testing on the above conclusions that are summarized in Figure 3.3, I will propose a new hypothesis that relates characteristics of pluripotent networks with certain motif distributions and test whether it can hold to be true in small random networks.

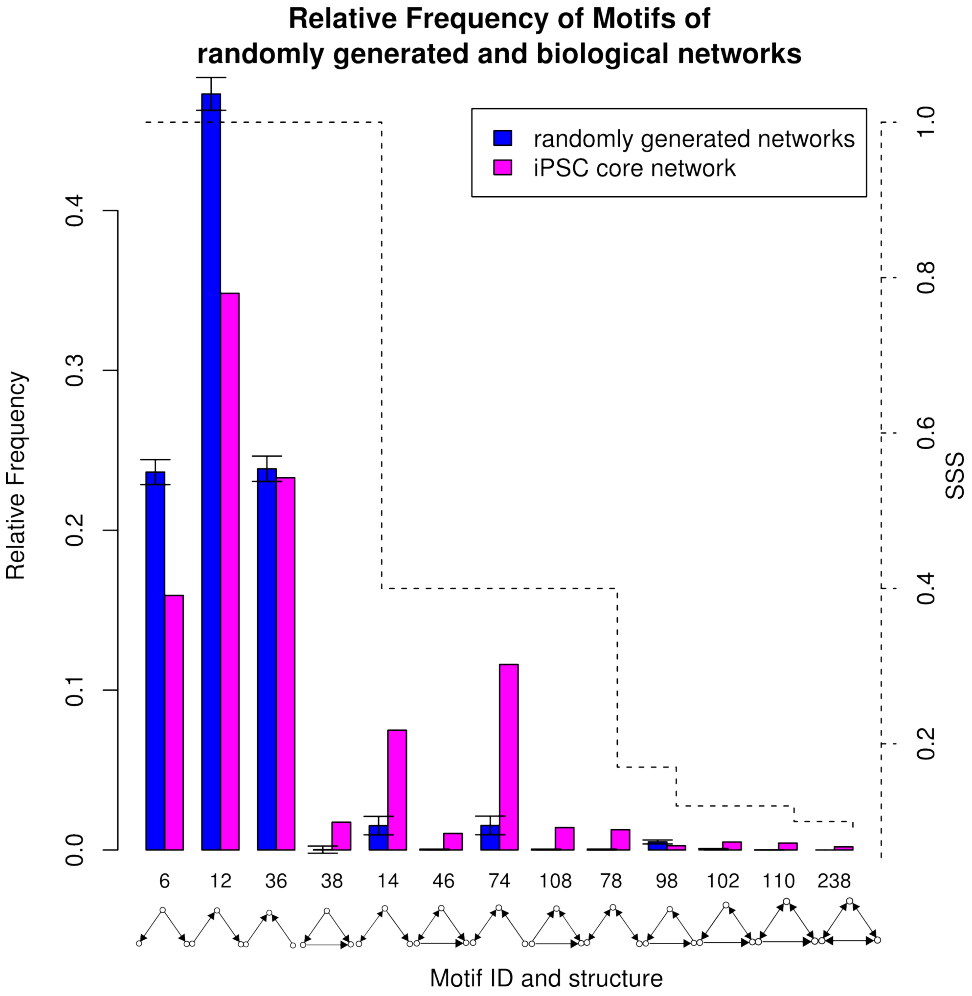

Figure 3.3: Motif Frequencies in the iPSC network in comparison to random Boolean networks and Relation to the structural stability score (SSS)

The relative frequency of Motif occurrence in the iPSC network and the mean in 100,000 RBNs is shown. Error bars represent standard deviations. At the bottom of the graph the motif IDs as enlisted in the mFinder tool are shown together with their appearance. The dotted line indicates the SSS of the motifs. In order to show the motifs in a decreasing order with respect to their SSS, they had to be partly re-arranged which is why the motif IDs are not constantly increasing. It can be seen in general that structurally stable motifs on the left are significantly under-represented in the iPSC while motifs with an intermediate SSS of 0.4 and even with lower SSS are over-represented

3.2 Does a Certain Configuration of Stable and Instable Attractors of a Network Influence its Motif Distribution?

As mentioned above, it appears that a network involved in pluripotency and differentiation needs to be multi-stable (Macarthur et al., 2009), i.e. have multiple attractors in which the system can reside, one of these attractors corresponding to the pluripotent cells and others corresponding to different differentiated cell lineages. It has been shown that when master regulators of differentiation are involved in the pluripotency network, the system shows bi- or multi-stable behavior with stable steady states corresponding to the differentiated states and an unstable steady state corresponding to the pluripotent state, describing the system as highly unidirectional (MacArthur et al., 2008; Chickarmane and Peterson, 2008). In a nice work combining experimental and theoretical approaches, this instability of the pluripotent state was attributed to fluctuations in *NANOG* expression (Kalmar et al., 2009). Moreover, I have shown in my diploma thesis that in small core Boolean networks of pluripotency, the basin sizes of the pluripotency associated attractors in the state space are extremely small compared to the attractor basins of differentiation related steady states (unpublished results). In Boolean networks, the size of the basin of attraction in the state space, i.e. the number of states that it includes, is linearly correlated to the stability of the attractor in question (see Section 2.3.3 in the introduction), i.e. small basins of attraction account for lower stability. The general hypothesis that pluripotency associated attractors are less stable than differentiation associated attractors is also strongly supported by the fact that differentiation of pluripotent cells occurs spontaneously or can easily be induced via external factors such as BMP4, Activin hypoxia and takes only little time (Greber et al., 2008; Prado-Lopez et al., 2010; Kubo et al., 2004) while the reprogramming of differentiated cells to iPSCs is still experimentally challenging, has low efficiency and takes long reprogramming times (Takahashi and Yamanaka, 2006).

I have found in the last Section that 3-node motifs with a high structural stability score - that are suspected to convey stability to the attractors of a network - are significantly under-represented in the iPSC network in comparison to RBNs which share the same number of nodes and mean number of inputs while motifs with lower SSS are significantly over-represented. At the same time, as mentioned above, it was found in several independent studies, that pluripotency related networks appear to follow the concept of multistability, i.e. that they have multiple different attractors corresponding to the different developmental states. Moreover, it is a very sensible hypothesis that the pluripotency related attractor is less stable than the differentiation

related ones.

Therefore, I generated the hypothesis that the specific stabilities of the network's attractors and thus the structure of the state space of a corresponding Boolean model are strongly correlated with the motif distribution.

In order to test this hypothesis, I constructed constrained networks by filtering 150,000 randomly generated 10-nodes Boolean networks for specific characteristics. As shown in the introduction in Subsection 2.3.3, the basin size of an attractor in a Boolean network is linearly correlated to its stability. The filter criteria are enlisted in the following:

1. The network is only allowed to have point attractors
2. The network at least has 3 point attractors
3. Exactly one of the point attractors is less stable than the others reflected by a smaller size of its basin of attraction

It is now possible to understand why random Boolean networks were constructed in the first place because the filtering for the basin sizes requires a Boolean network. It is important to understand how the basin sizes were determined for the filtering. In a Boolean network with n nodes, the state space consists of 2^n states. In order to find out, what an average size for an attractor's basin size would be, I computed the expected value of the relative basin size in the network as:

$$E^{basin} = \frac{1}{N_{attr}} \tag{3.3}$$

where N_{attr} is the number of attractors inside the network in question. In order to find the expected value of the real basin size, one would have to multiply this relative number with the number of states 2^n. For clarity, it should be mentioned at this point that this number has nothing to do with the scaling of attractor number and length with system size (Kaufman et al., 2005; Drossel et al., 2005) but it is the expected average attractor length, when the number of attractors is already known.

Next, this value is compared to the real relative size of the basin of the attractor in question:

$$R_{attr}^{basin} = \frac{S_{attr}}{2^n} \tag{3.4}$$

If $R_{attr}^{basin} < E^{basin}$, the basin of attraction of the attractor in question can be considered as small. However, in order to test my hypothesis, I constructed

networks where this constraint is constantly intensified, i.e. the condition $R_{basin} < E_{basin}$ is not enough anymore, but rather:

$$R_{basin} < \frac{1}{2^m} * E_{basin} \tag{3.5}$$

with $m \in \{1..5\}$ a natural number which leaves us with 5 sets of networks in which the smallest basins of attraction are at least twice and at most $2^5 = 32$ times smaller than the expected average basin size and the second smallest basin has at least the size of the expected basin size. Therefore, in these sets of networks the smallest attractor decreases in stability when m increases. In the following, these 5 sets of networks will be called the *basin filtered networks* or *basin size filtered networks*. Due to computational constraints in the search of attractors for a network, only networks up to 29 species are supported in BoolNet. For the filtering of the networks, however, the search of attractors is inevitable which is a strongly limiting step with regard to computational time and power. In order to keep computational times reasonable and have consistent results, I generated 150,000 random Boolean networks of 10 nodes with an average number of neighbors of 3.90625 which equals that of the iPSC network. The number of networks found in every set is shown in Table 3.1 together with the total number of random networks of 10 nodes and 125 nodes.

Table 3.1: Number of Networks in the Different Sets

Shown are the number of networks in the set of random 125-nodes networks, in the random 10-nodes networks and in the networks that were filtered for the different basin sizes based upon Equation 3.5. The factor $m \in \{1..5\}$ from that equation appears in the filtered networks' name by appending the number 2^m to the word *Basin*

Network Set Name	**Number of Networks**
Random 125-nodes networks	100,000
Random 10-nodes networks	150,000
Basin2	234
Basin4	394
Basin8	354
Basin16	234
Basin32	127

The hypothesis is that the filtered networks and especially the ones with smaller basin sizes, favor a motif distribution that is more similar to the one of the iPSC network than it is to the one of random networks. This hypothesis will be tested in the following by progressively applying the statistical tests from the decision tree (Figure 2.2) described in Subsection 2.3.1.

I will start with the Shapiro-Wilk tests for normality for the 100,000 125-nodes networks and the 150,000 10-nodes networks, proceed to the variance tests, then to the 2-sample comparisons of different combinations and finally end with the 1-sample t-tests comparing the mean of distributions with the relative frequency of the iPSC network.

The Shapiro-Wilk test is limited to samples of size n=5000. However, I have samples with up to 150,000 values. In fact, when performing a few test runs for different samples of the 125-node network's motif distributions, it gives very different results sometimes rejecting sometimes accepting the null hypothesis at a significance level $\alpha = 0.05$. This is partly reflected in Figure 3.2, where some of the low ID motifs at least graphically appear to have normal distributions while the majority clearly have non-normal or skewed distributions. However, the strongly different results were also found inside the distribution of one motif only. Therefore, the Shapiro-Wilk test with samples of size 5000 out of the 100,000 networks, was performed 10,000 times for every motif. It is then possible to calculate a mean and a median of the resulting p-values of the test applied to these samples. As could be expected by the very different results in the test runs, the mean and the median show strong discrepancies pointing at a non-normal distribution of the p-values of the 10,000 tests. In fact, I found that the median always had significantly lower values than the mean indicating that most p-values are low and close to the median and very few have very high values shifting the mean to higher values. This was confirmed by plotting histograms of the p-values (results not shown). Since for these skewed, strongly non-normal distributions, the median is the better choice to show central tendencies, it was chosen to evaluate the ensemble of Shapiro-Wilk tests. Moreover, the mean of the p-values is not the important measure in the case of hypothesis testing. It is more important how often the p-value of the Shapiro-Wilk is below or above the threshold of $\alpha = 0.05$ thus respectively rejecting or accepting the null hypothesis at that significance level. However, the median of the sample of p-values on its own does not have enough power to accept or reject the null hypothesis. Therefore, I constructed a confidence interval of the median following the definitions by Conover (1980) and Bland (1995). For large samples, such as the one we generated by running the Shapiro-Wilk test 10,000 times, they define the number of values of the sample that are lower than the q-quantile as an observation of a Binomial distribution with parameters n and q, respectively representing the sample size and the quantile number which is 0.5 for the median. The confidence interval can then be calculated with the following equations:

$$upper = nq + 1.96 * \sqrt{n * q(1-q)}$$
$$lower = nq - 1.96 * \sqrt{n * q(1-q)}$$

After setting $q = 0.5$, we get the Equations 3.6 that are used to derive the ranked values of the confidence interval:

$$upper = \frac{n}{2} + 1.96 * \frac{\sqrt{n}}{2}$$
$$lower = \frac{n}{2} - 1.96 * \frac{\sqrt{n}}{2} \tag{3.6}$$

It is important to notice that the 2 values calculated from Equations 3.6, are used to determine the upper and lower ranked values in the sample which then represent the extremes of the confidence interval. Since the values are not integers in the majority of cases, they are rounded, the *lower* to the next integer below, the *upper* to the next integer above. The sample is then ordered and the values inside the ordered sample at the place of the lower and upper integer ranks make up the extreme values of confidence interval. Therefore, opposite to confidence intervals for the mean, the confidence interval of the median is not symmetrical, i.e. it can be more extended in one direction than in the other. These Shapiro-Wilk sampling test results for the 125-nodes network are represented in Table 3.2.

Except for motif 5 with the mFinder motif ID 38 (see Figures 2.3 and 3.2), which has a rounded median of p-values of 0.075 ± 0.005 for the Shapiro-Wilk test, we could reject the null hypothesis that samples come out of normal distributions at a significance level $\alpha = 0.05$ for all other motif distributions. Although graphically motifs 6, 12 and 36 strongly appear normally distributed, the hypothesis could not be verified in the test. It should be stated, that the Shapiro-Wilk test is extremely sensitive to outliers in big samples. However, it still constitutes the normality test with the highest power in comparison to the Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests (Razali and Wah, 2011).

In a similar manner, I ran the Shapiro-Wilk test for normality 1000 times with different samples of size 5000 out of the 150,000 random networks with 10 nodes for every one of the 13 motifs. In this case, the null hypothesis could always be rejected without a doubt with mean p-values $< 2.2 * 10^{-16}$, i.e. very low probabilities that the samples are drawn out of normal distributions

Table 3.2: Shapiro-Wilk Sampling Results for 125-nodes Networks

We represent the mean, standard deviation (sd), standard error (stderr), symmetric confidence interval of the mean (mean C.I.), the median, and the lower and upper bounds of the asymmetrical confidence interval of the median (Explanation see plain text)

Motif	mean	sd	stderr	mean C.I.	median	lower	upper
1	0.021	0.071	0.001	0.000	0.001	0.000	0.000
2	0.014	0.058	0.001	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.016	0.062	0.001	0.000	0.000	0.000	0.000
5	0.19	0.244	0.002	0.000	0.075	0.005	0.005
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.001	0.007	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.000	0.000	0.000	0.000	0.000
13	NA	NA	NA	NA	NA	NA	NA

(results not shown). In a first step, we can now try to assess whether the 10-nodes and 125-nodes RBNs which were constructed using the same average number of neighbors, show similar motif distributions.

Since the motif distribution samples from the 10-nodes networks are all non-normally distributed, the Levene test was carried out between the 150,000 random 10-node networks and the 100,000 125-node networks for every motif in order to assess whether the 2 samples in each case (for every motif) have equal variances. The null hypothesis could be rejected with a very high confidence (p-value of the Levene test $< 2.2 * 10^{-16}$) for every motif (results not shown). This is to say that the probability is infinitely low that the two samples come from distributions with equal variances. In order to compare the two distributions, we must thus apply the Kolmogorov-Smirnov test which does not make any assumptions concerning the nature of the original distributions (Kolmogorov, 1933; Smirnov, 1948). For every motif, the Kolmogorov-Smirnov test similarly yielded a p-value $< 2.2 * 10^{-16}$ and although destined for samples of same variance, the Wilcoxon-Mann-Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947) was also run on the two samples yielding the same result as the Kolmogorov-Smirnov test. The main reason why the Wilcoxon test was run as well is that the Kolmogorov-Smirnov test is very sensitive to so-called ties, i.e. equal values in the sample. It should be noted at this point, that although the Mann-Whitney (Wilcoxon) test analyzes the null hypothesis that the two distribution functions don't differ by a location shift, this implies also the null hypothesis that the means

are the same against the alternative that they are not and the null hypothesis that the medians are the same against the alternative that they are not.

Taken together, the results above clearly show that the motif distributions in the 150,000 random 10-node and 100,000 random 125-node networks, although having the same average number of neighbors, are in fact topologically very different from each other. This might be due to scaling effects, as the mean number of appearances of subgraphs, and thus also network motifs scales with the network size and the mean connectivity (which should be the same in both networks) (Itzkovitz and Alon, 2005). Since all subgraphs scale in this way and since I normalized motif occurrences to the total amount of motifs in the network, the system size should not affect this measure. However, it is still possible that a 10-nodes network is too small to apply the scaling argument above and that due to finite-size effects the number of subgraphs does not scale with the complete network size N but rather with N^{n-g}, where n is the number of nodes of the subgraph in question and g is its number of edges. This scaling is called Erdős-Rényi scaling (Bollobas, 1985; Itzkovitz et al., 2003). Due to this fact, the interpretation of the following results should be treated with care. In fact, for this reason, the basin filtered 10-nodes networks will not be compared to the random 125-nodes networks but only to the random 10-nodes networks in order to draw conclusions.

We will now proceed to the testing of our hypothesis that in the sets of networks which are filtered for their basin size out of the 150,000 RBNs with 10 nodes and with an average number of neighbors of 3.90625, the mean of the relative motif frequency distribution resembles more the value of the relative motif frequency in the iPSC network than in the random 10-nodes networks. In order to test this, I will first carry out a pairwise comparison of the basin filtered networks with the 150,000 10-nodes random networks from which they were filtered in order to show, that they effectively constitute different samples. Since for every sample combination, the decision tree from Figure 2.2 needs to be taken into consideration, I designed a series of tests using R scripts that carry out the tests for the different samples and motifs automatically.

The tests show that except for motif 5 (ID 38), which has a very low mean relative frequency in the 100,000 random 125-node networks of around 0.015, the null hypothesis of the Kolmogorov-Smirnov and the Wilcoxon or the t-test could always be rejected for all sample combinations, proving that the basin filtered networks, although filtered out of the 150,000 10-nodes RBNs, show a significantly different distribution of the network motifs' relative frequencies. These test results can also be graphically double-checked in the summarizing boxplots in Figure 3.4.

For the basin filtered network samples, the results are a little more complex.

While in the majority of cases, the null hypothesis can be rejected here as well, there are combinations of basin sizes and motifs whose relative frequency distribution can actually be regarded as having a mean corresponding to the relative motif frequency of the iPSC network motif distribution. It was thus found that for motifs with numbers 3, 5, 7, 8, 9, 10, 11 and 13 (or IDs 14, 38, 74, 78, 98, 102, 108 and 238 respectively as can be deduced from Figure 2.3), the null hypothesis could always be rejected for every basin filtered network. For motif 12 (ID 110) on the contrary, the null hypothesis was accepted for every basin filtered network. Except for motif 12, the null hypothesis is also accepted for the basin filtered networks with $m = 5$ for motifs 1 (ID 6), 4 (ID 36) and 6 (ID 46), for basin filtered networks with $m = 4$ for motifs 2 (ID 12) and 4 (ID 36) and for the basin filtered network with $m = 2$ for motif 2 (ID 36).

Motif 12 (ID 110) has the special characteristic that all basin filtered networks seem to have the same mean as the relative frequency of this motif in the iPSC network. It should be said that the majority of values in the samples with basins that are smaller than average by the factor 2,4,8,16,32 are 0 for Motif 12. The number of values that are not 0 are respectively 22, 42, 37, 23, 14 for the different samples respectively which are quite low percentages around 10% of the total number of networks in this sample (see Table 3.1). This means that in most of the filtered networks, motif 12 does not appear while in the random 10-nodes network it appears significantly more often. In the iPSC as well as in the random 125-nodes networks it has very low relative frequencies.

It is interesting to see that whenever the null hypothesis is accepted, it is mostly for motifs with a high SSS while it is nearly always rejected for motifs with lower SSS except for motif 12, the latter possibly being due to the very low number of occurrences that are not 0 in both distributions. However, the results are diffuse throughout the basin filtered networks with different values for m. It could thus be that a general tendency is distinguishable that networks with one basin of attraction being at least 2 times smaller than average partly have a similar mean relative frequency of their motifs as the iPSC network but it cannot be said with certainty that a direct correlation exists.

The complete test results of the last paragraphs can be summarized by the boxplots in Figure 3.4.

3.3 Summary and Discussion

In this Chapter, I have first shown that a network involved in processes of induced pluripotent stem cells has a significantly different motif distribu-

tions than homogeneous random networks. In a general tendency, it can be said that motifs with high SSS which show stable behavior and are hypothesized to enhance the stability of the system are under-represented in the iPSC network while motifs with a lower SSS are over-represented. It was hypothesized that this could partly be due to a decreased stability of the attractors of a corresponding pluripotency model that needs to have dynamical plasticity in order to differentiate into different cell lineages upon external triggers and partly to the difference in distribution between homogeneous and scale-free networks.

In the second part, I analyzed whether the stability condition for small Boolean networks that one attractor (corresponding to the pluripotent state) is less stable than the other attractors is enough to account for a similar motif distribution observed in the iPSC network. However, this could not be verified, on the contrary it seems that the condition is not enough. Although the general trend of the mean of relative motif frequencies of the basin filtered networks is overall slightly closer to the one of the iPSC network than the random 10-nodes networks (see the Boxplots in Figure 3.4), it is difficult to speak of a direct correlation. This is also due to the strong spreading of relative frequencies in the 10-nodes random and filtered basin networks in comparison to the 125-nodes networks. It is possible that the constraints taken as a basis for the filtering are partly inducing the new topological trends and becoming more similar to the iPSC network from motif abundance point of view. Due to computational and software limits concerning the attractor search and stability analysis, the hypothesis could only be tested for small RBNs of 10 nodes which were shown to also have strongly different motif distributions than the random 125-nodes networks with the same average number of neighbors. This could be explained with finite-size effects for small networks where the scaling of subgraph abundance is not proportional to the number of nodes in the network but to the number of nodes subtracted by the number of edges involved in the subgraph which is called Erdős-Rényi scaling. Due to this constraint, it was only possible to draw conclusions as to how the filtered basin networks behave in comparison to the random 10-node networks but not in comparison to the 125-nodes networks which could only be compared between them. In order to gain more certainty on this matter, more complex analyses would have to be carried out with bigger networks which, however, is computationally limited by the attractor search and perturbation algorithms.

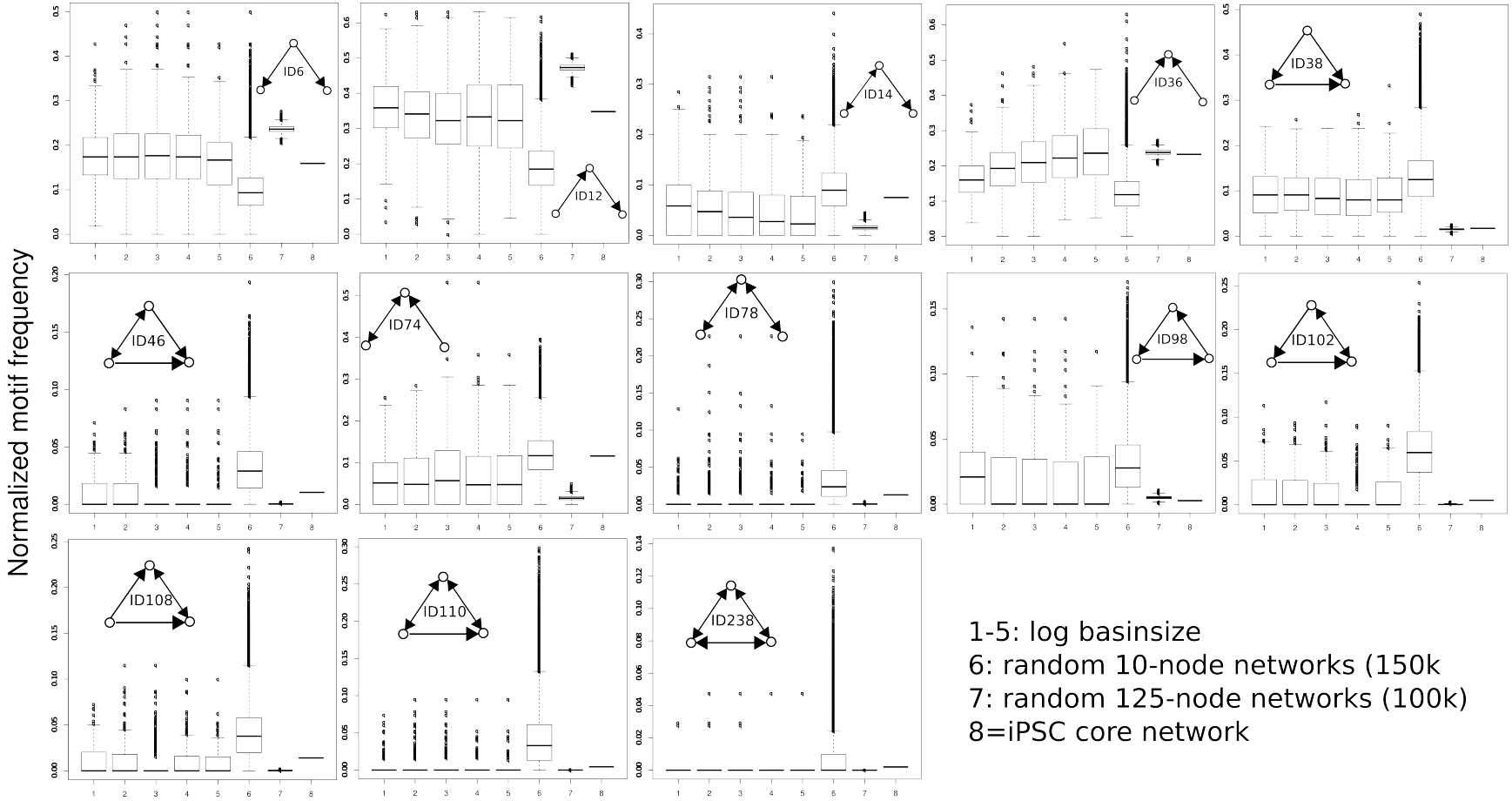

Figure 3.4: Summarizing Boxplots

The normalized motif frequency is shown on the y-axis against different experiments on the x-axis (see legend in the lower right). The first five boxes of every plot represent the sets of random 10-node networks filtered for basin sizes that are smaller than average by the factor 2^m with m corresponding to the placement 1-5 of the boxes. The boxes six and seven represent the random 150,000 10-node and 100,000 125-node networks respectively. The last part of each plot corresponds to the relative motif frequency of the iPSC core network.

4 Training of a Boolean Model Against Reprogramming Data Unveils New Insights into the First Steps of Reprogramming

4.1 A Confident Transcriptional Interaction Network: Automated Literature Mining, Expert Curation and Data Enrichment

The network motif analysis done in the last Chapter, was performed on a big automated literature mining network, created using the Genomatix Pathway System tool (GePS) developed by Genomatix on the basis of LitInspector (Frisch et al., 2009). For slightly more details on the algorithm and functions of this tool, please consult Section 2.2.2.

The network in question consists of the core regulatory circuitry of pluripotency of OCT4, SOX2 and NANOG (Boyer et al., 2005) embedded into a more extended gene regulatory network (GRN), enriched with protein protein interactions (PPIs), epigenetic regulations and small molecules to only name a few. When looking at Section 1.2, which only treats of signaling pathways, transcription and epigenetics as 3 out of many different regulatory processes, one can approximately estimate what it means to build a model that will at least partly reflect biological reality. To be able to gain insights into the functioning of this big machinery, the goal of the theoretical biophysicist is to start with a small network and steadily increase its size or start with a big automated network and steadily decrease its size in

order to attain a meaningful extent. It is necessary to keep in mind that a bigger network will only be advantageous if there is data available for the species and mechanisms inside the network in question, which requires a close cooperation between theoretician and experimentalist.

The aim of this Chapter is the construction of a reasonably sized transcriptional interaction network involved in pluripotency, the translation of this network into a Boolean model and the subsequent training and analysis of this model to microarray expression profile data of early reprogramming in order to unravel the important mechanisms in the first 96 hours of the process.

The network stored in the Genomatix pathway database as the iPSC core version 2 pathway which was used in Chapter 3 contains 125 genes, proteins and small molecules that have been known to play a role in induced pluripotent stem cells. In order to reduce the size of the network and focus on the most important parts, it is enriched with the gene expression profiling microarray from the experiments described in Subsections 2.1.1 and 2.1.2 and explicitly shown in the appendix in Section A.1.

We are dealing with microarray expression profiles from different transduction experiments, i.e. perturbation data measured upon different conditions. In our data, the conditions are set by the different retroviral vectors with which the human fibroblast cells were transduced: They either carry the gene encoding one of the 4 transcription factors of the Yamanaka reprogramming cocktail POU5F1, SOX2, KLF4, c-MYC (Takahashi and Yamanaka, 2006) or a combination of the first 3, called *3TF* or all 4 called *4TF*, summing up to 6 different conditions. The last two conditions *3TF* and *4TF* are the reprogramming conditions that upon a certain time span lead to iPSCs. The unperturbed condition is called *FIB*, because it is nothing else than the expression profile of fibroblasts. There is also a perturbed measurement of fibroblasts transduced with a retroviral vector carrying the *GFP* gene. The exact values of the raw microarray gene expression profiling are shown in the appendix in Section A.1.

In first place, these microarray data will now be used for the filtering of the above mentioned network. In fact, I will only consider species with a significant detection p-value < 0.01 that are differentially expressed (signal intensity ratio $\frac{factor}{mock} > 1.5$ for up-regulation or < 0.67 for down-regulation). For more details on the data processing see Subsection 2.1.2. After applying this filtering, a network of 39 nodes emerges whose characteristics will partly be analyzed in the following.

One striking feature of the filtered network is that *NANOG*, a gene known to play a prominent role in ESCs (Mitsui et al., 2003), is not a part of it since it is not differentially expressed for any of the measurements. Since *NANOG*

is exclusive for pluripotent stem cells and is only expressed at late stages of reprogramming (Brambrink et al., 2008; Silva et al., 2009), it might be that it doesn't play any role in early reprogramming. It was therefore left out of the network due to a lack of potential information. SOX2 and POU5F1 had insufficient p-values for at least one measurement but not for all of the 7 measurements (*FIB/GFP* and the six combinations). However, as it was already shown by Boyer et al. (2005), these 2 factors play an important role in pluripotency together with the NANOG gene which has intensely been used in the field of reprogramming to mark the final transition to iPSCs. Therefore, SOX2 and POU5F1 were re-introduced into the network. With the process described so far, a network of 41 differentially expressed genes and 295 automated interactions between those genes is created. This literature mining approach is nice to get an idea about the network size, topology and the species and mechanisms involved. However, it is not possible to completely rely on it for data integration and modeling. This is mainly due to two reasons:

1. The literature mining approach generates many false positive interactions

2. It generates all possible interactions between two genes or gene products, i.e. different PPIs, transcriptional interactions, epigenetic interactions, etc. although one may only be interested in one special type of interaction.

Apart from creating a reliable interaction network for processes that take place in reprogramming, my goal was also to continue working with the network to fit a model to the microarray data. Since the latter are just a way to measure the quantity of mRNA, it is reasonable to focus mainly on transcriptional interactions in the network or on other processes that directly influence the quantity of mRNA production without other mechanisms lying in between. If we were to include all PPIs as well, this would mean, we would be looking on a different scale and our data would be insufficient to characterize this kind of interaction. To eliminate interactions that don't influence mRNA quantity directly, I scanned through the literature data of all the putative interactions, looking mainly for transcriptional links and curating the network in a detailed manner. This very exhaustive work yielded a highly confident mainly transcriptional interaction network with 26 genes and 75 transcriptional interactions. The interactions together with literature evidence have been summarized in Table 4.1. The raw microarray data at the basis of this approach is shown in Table A.1. As such, the transcriptional interaction network regulating pluripotency is unique in literature and can be used for further modeling approaches, knockout experiments and as a stand alone network, can help to gain insights into mechanistic features in the future.

Table 4.1: Big Curated Pluripotency Network

The gene regulatory network is represented in tabular form: The first three columns resemble the structure of a SIF file with the effecting transcription factor in the left column, an interaction (1 for activation, -1 for inhibition) in the middle and the regulated gene (or its encoded transcription factor) on the right. In the last column, I included references and explanations supporting the interaction

Transcription Factor	Interaction	Regulated Gene	Reference
POU5F1	1	CARM1	Wu et al. (2009)
ID2	-1	CCND1	protein level, Tokuriki et al. (2009)
PARP1	1	KLF4	Gao et al. (2009)
PARP1	1	MYC	Carbone et al. (2008)
SP1	1	TGFBR2	Jennings et al. (2001); Huang et al. (2005); Ammanamanchi et al. (1998); Periyasamy et al. (2000)
TBX3	1	POU5F1	TBX3 maintains pluripotency, binds to POU5F1 promoter and is known to directly upregulate NANOG. The interaction is thus a sensible speculation Han et al. (2010); Niwa et al. (2009)
KLF4	-1	CCND1	Shimizu et al. (2010); Shie et al. (2000)
ID2	1	SP1	Partial evidence for expression-inducing SP1 in ID2 promoter. Kurabayashi et al. (1994)
PARP1	1	HIF1A	Elser et al. (2008)
ID2	-1	MYC	Torres et al. (2009); Rodríguez et al. (2006)
MYC	1	ID2	Coma et al. (2010) (plus Matinspector binding site from Genomatix)
POU5F1	1	FGF2	Greber et al. (2007b) plus ChIP-on-chip data by Boyer et al. (2005)
IRS1	1	CCND1	Sun and Baserga (2008); Wu et al. (2008)
STAT3	1	MYC	Kiuchi et al. (1999); Bowman et al. (2001)
HIF1A	-1	CCND1	Wen et al. (2010)
POU5F1	1	SOX2	Boyer et al. (2005)
STAT3	1	KLF4	Bourillot et al. (2009) only in mouse
IRS1	1	STAT3	Sun and Baserga (2008)
SP1	1	FGFR1	Seyed and Dimario (2007)
EPAS1	1	POU5F1	Covello et al. (2006)

STAT3	1	EPAS1	Korgaonkar et al. (2008)
STAT3	1	CDK4	Radaeva et al. (2004)
SMAD3	1	GREM1	Zode et al. (2009) (and Matinspector binding site from Genomatix)
SP1	1	EPAS1	Wada et al. (2006)
PARP1	1	SOX2	Higher SOX2 protein stability through polyadenylation. Gao et al. (2009)
SP1	1	PARP1	Laniel et al. (2004); Zaniolo et al. (2007)
KLF4	1	MYC	Liu et al. (2008), Boyer et al. (2005), Kim et al. (2008)
KLF4	1	POU5F1	
KLF4	1	KLF4	
POU5F1	1	POU5F1	
SOX2	1	SOX2	
SOX2	1	POU5F1	
HIF1A	-1	PTPRU	ten Freyhaus et al. (2011)
POU5F1	-1	ID2	Babaie et al. (2007)
POU5F1	1	STAT3	Boyer et al. (2005)
STAT3	1	POU5F1	Do et al. (2013); Kim et al. (2013); Som et al. (2010)
SP1	1	CDK6	Cram et al. (2001); Firestone and Bjeldanes (2003)
SP1	1	CCND1	Kitazawa et al. (1999); Nagata et al. (2001); Huesca et al. (2009)
CCND1	-1	SP1	Shao and Robbins (1995); Adnane et al. (1999)
SP1	1	MYC	Majello et al. (1995); Geltinger et al. (1996)
PTPN11	1	CCND1	PTPN11 phosphorylates ANGII and thereby enhances CCND1 epxression (indirect transcriptional interaction) citepGuillemot2000
EPAS1	1	SOX2	Moreno-Manzano et al. (2010)
KLF4	-1	GSK3B	Effectively there is a binding site for KLF4 in the GSK3B promoter (Boyer et al., 2005) and GSK3B needs to be down-regulated in ESCs (to stop phosphorylation and thus degradation of beta Catenin). Therefore this interaction represents a reasonable hypothesis

KLF4	-1	ID3	Nickenig et al. (2002)
KLF4 SP1	-1 1	SP1 SP1	Kanai et al. (2006)
SP1	1	KLF4	Mahatan et al. (1999)
SP1	1	IRS1	Panno et al. (2006)
STAT3	1	SOX2	Foshay and Gallicano (2008)
STAT3	1	HIF1A	Marzec et al. (2011); Xu et al. (2005)
CARM1	1	SOX2	Wu et al. (2009); Torres-Padilla et al. (2007)
SOX2	1	CCND1	Chen et al. (2008)
MYC	1	CDK4	Obaya et al. (1999); Hermeking et al. (2000)
SP1	1	CDK4	Willoughby et al. (2009)
CDK4	1	SP1	Through protein-protein-interaction (PPI) with SP1 (which increases its own production) (Tapias et al., 2008)
CCND1	-1	TGFBR2	Zhang et al. (1997); Okamoto et al. (1994)
SP1	1	HIF1A	Koshikawa et al. (2009); Kim and Park (2010)
SMAD3	-1	CDK4	Matinspector binding site from Genomatix and Wolfraim et al. (2004) or indirectly over CDK inhibitor p15 which is regulated by SMAD3 (Matsuura et al., 2004)
HIF1A	1	ID2	Löfstedt et al. (2004)
MYC	1	CCND1	Swarbrick et al. (2005); Yu et al. (2005) and binding site in CCND1 promoter (Boyer et al., 2005) and Matinspector from Genomatix
HIF1A	1	FGF2	Black et al. (2008)
STAT3	1	FGF2	xin Xie et al. (2006)
TLE1	-1	MYC	Sierra et al. (2006)
STAT3	1	CCND1	Turkson and Jove (2000)
TLE1	-1	CCND1	Fraga et al. (2008)
SMAD3	-1	MYC	Frederick et al. (2004)
IRS1	1	MYC	Wu et al. (2008)

4.2 Integrating Prior Knowledge Networks and Perturbation Data to Optimize a Boolean Model

In the Section above, I have outlined how a confident transcriptional interaction network involved in early reprogramming was established and curated. In Subsections 2.1.1 and 2.1.2, the experimental details for the generation of the microarray gene expression profiling data of early reprogramming experiments (shown in the appendix in Section A.1 are described as well as their analysis and processing. I will now describe how these two sources of information - the literature network and the experimental data - are brought together in the framework of the CellNetOptimizer (CNO) package for R (Terfve et al., 2012) to optimize a Boolean logic model, the advantages and disadvantages and the results and insights gained using the method. It should be said in advance that in the following the network as well as the data will gradually be reduced or extended at certain parts and normalized (for the data) and thus will progressively change their appearance until they are compatible with biological requirements, underlying mechanisms and the software tool to start the optimization. I will progressively enumerate reasons for discarding certain genes, extending the network by certain others or differentiating between endogenous and exogenous versions of genes. I will moreover explain how the software works to bring together data and network. It is important to understand these steps in order to be able to retrace the final network and data set as well as the optimization and its results.

Although a detection p-value of 0.01 was fixed in order to only choose differentially expressed genes involved in the reprogramming process as explained above in Section 4.1, a few sensible exceptions have been made in order to get the best possible data set for optimization. These exceptions only concern two species that are known to be involved in the reprogramming process and that constitute two of the main master regulators of pluripotency: *POU5F1* and *SOX2*. In fact, the endogenous *POU5F1* gene expression has p-values above 0.01 for the *FIB* and *GFP* measurements as well as for the *SOX2*, *KLF4* and *MYC* assays, while the SOX2 gene expression has p-values that are above 0.8 for every assay. These high p-values are always associated with very low gene expression values that don't significantly differ from the background measurement and can thus be associated with a lack of expression. Therefore, in the cases, where the p-values are high and thus gene expression is very low or undetectable, the data for the condition and time point in question was set to 0 after rescaling of the data. This allows us to include SOX2 and POU5F1 into the model and data and thus into the optimization process in a sensible way. It would have been possible to leave out or set to NA the measurements associated with insufficient detection p-values. How-

ever, all optimization yielded better results for the case where SOX2 and POU5F1 measurements are assigned the sensible values and since expression of these two genes is specific to stem cells, this assumption is justified.

The data is processed and normalized using the rescaling method as outlined in Subsections 2.2.5 and 2.1.2 and loaded as a MIDAS (Minimum Information for Data Analysis in Systems Biology) file (described in detail in Saez-Rodriguez et al. (2008)). Originally, the CNO package was designed for signaling pathways and the data loaded in the MIDAS file were thus related to modification states of proteins, mostly phosphorylations that activate or deactivate the signaling protein in question. However, in the same way it is possible to describe gene regulatory networks (GRNs). The active or inactive proteins then correspond to expressed or unexpressed genes and the interactions between species that before were protein protein interactions (PPIs) can now be interpreted as transcriptional interactions between one transcription factor and a gene.

In fact, it is possible to either normalize the data against the *FIB* or the *GFP* conditions mentioned above. When normalizing against the latter, it is possible to account for the differential gene expressions induced by the retroviral vector integration triggering the virus response of the cell which is probably not accounted for in the pluripotency network to be optimized. It is thus reasonable to normalize against the *GFP* control. Since I will be using rescaling as a means to normalize data (as explained further below), this means that the *GFP* data point will be taken as time point 0 in the rescaling method. Interestingly, in the course of my research, I have discovered that when taking the *FIB* control as data point at time 0, the optimization results are slightly worse than when including the *GFP* data which supports the use of the latter. The better results gained by optimizing against the *GFP* control are probably due to the fact, that the effects of the virus response can not fully be accounted for by our network model that mainly includes genes that play a role in iPSCs. Therefore, the genes regulating the virus response of the cell are unlikely to be included in this set.

Following the data processing, the prior knowledge network (PKN) which in theory would be the network generated by literature mining, data integration and curation in Section 4.1 above, is loaded as a SIF file describing directed interactions between species (just as in Table 4.1). This PKN is expanded into all possible Boolean models as described in Subsection 2.2.5. As we will see in the following, the network will need to be further modified.

In order to be able to understand the computational effort and the sense of optimizing a network of this size as the network in Table 4.1, I will now do a quick estimation of the state space that is searched by the tool: We are dealing with a highly interconnected network. A network analysis with Cytoscape yields the network statistics displayed in figure 4.1.

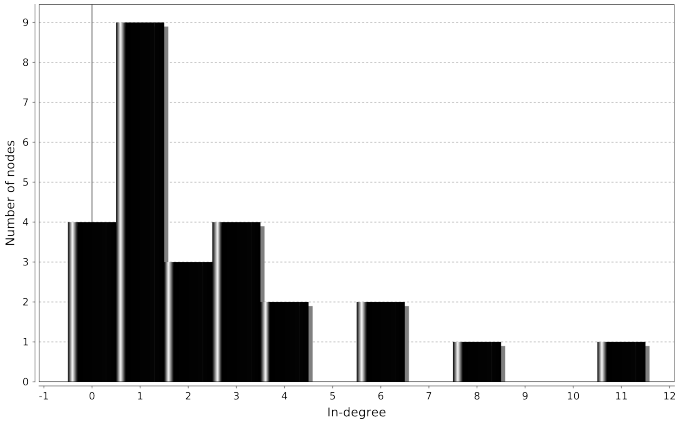

Figure 4.1: In-degree distribution of the big pluripotency network from Table 4.1.

As explained in 2.3.3, the possible Boolean functions for one species strongly depend on the number of inputs to the node and increase by 2^{2^n} (if there is no knowledge at all about the quality of the interaction) where n is the number of input nodes. From fig. 4.1 one can quickly see that this means that if we only consider the node with the 11 inputs, this means there are 2^{2048} possible Boolean functions. With the approximation $2^{10} = 1024 \approx 10^3$, this yields a total of more than 10^{200} possible combinations only for one node. However, with the CellNetOptimizer package, this optimization state space that is valid for completely random interactions will strongly be simplified by introducing the prior knowledge about the network. Introducing the concept of activations and inhibitions eliminates a wealth of the possible input functions for each node: an activator will always change the state of its target from 0 to 1 or leave it at 1 while an inhibitor will always change the state of its target from 1 to 0 or leave it at 0. The exact mechanism of the Boolean logic gate expansion and optimization is outlined in Subsection 2.2.5. Nonetheless, although a lot of possibilities are eliminated the number of bits in the bit string that needs to be optimized for the big network is 975. Since every bit can take two values, 0 or 1 corresponding to presence or absence of the bit and thus of the corresponding logic gate, the optimization state space consists of $2^{975} \approx 10^{300}$ possible functions. It is very difficult to estimate how well an algorithm could possibly perform in order to find a reasonable solution in such an extended optimization space. However, it is for that reason that a gene cannot have more than 6 inputs (i.e. $2^6 = 64$ Boolean functions) or the software will encounter memory problems.

Due to the size of the optimization state space, computational limits and the fact, that part of the network still consists of interactions that are not 100% sure in literature or not purely transcriptional, I further reduced the network.

I first eliminated CARM1 out of the network, an epigenetic regulator that is already known to play an important role in the reprogramming process (Wu et al., 1999). Since epigenetic regulation partly takes place at a different time scale than transcriptional processes, it is very difficult to take these processes into account in a model that is only fed with gene expression profiling microarray data, i.e. data that in fact measures the mRNA concentration. In Chapter 5 I will focus more on the modeling of epigenetic processes together with other layers of regulation in a model of reprogramming and differentiation (Flöttmann, Scharp, and Klipp, 2012). Furthermore, interactions that were not highly confident from both literature and ChIP-on-Chip binding data were deleted progressively. The reduction yielded a network containing 18 endogenous genes plus the 4 retroviral genes (whose genuine name has been extended by the suffix *ext* in my network) and 53 interactions between them. However, a few modifications still need to be done.

Apart from the interactions that were drawn from literature analysis, there are a few key features of the network that have to be taken into consideration as well when trying to combine it with the microarray expression data. When taking a closer look at the microarray data, one can realize two important things:

1. The exogenous viral gene transcripts differ in their 5' UTR (untranslated region) that is especially important for the hybridization onto the microarray, i.e. only endogenous transcripts will be hybridized onto the chip and the retroviral gene expression will not be measured by the microarray.

2. The endogenous pluripotency related genes SOX2, POU5F1, KLF4 and c-MYC show only very low expression if any because they are in parts epigenetically masked or repressed by other transcription factors since they only show their highest expression in ESCs or iPSCs.

Considering these two facts, it is reasonable to change the topology of the PKN in the following way: I modified the interaction rules for the 4 endogenous equivalents of the retroviral pluripotency genes so that they only have incoming, no outgoing edges anymore. In fact, since the exogenous, non-measured genes are much higher expressed due to the specific promoters on the plasmid vectors, the main downstream effects will be due to those exogenous species. One of their downstream targets are as well the endogenous pluripotency genes whose expression is then measured on the microarray. It is this latter point of the network construction that eliminates the autoregulatory self-loops of the pluripotency genes that are known from Boyer et al. (2005). However, this reasoning is only valid for those measurements where the retroviral gene in question is included. For example, for the measurement where the viral POU5F1 gene is transduced, the other 3 exogenous genes are not present, i.e. there won't be any effect of the exogenous genes

for KLF4 and c-MYC for example on downstream targets. However, endogenous KLF4 and c-MYC still show expression and could act in these cases where they are not over-ruled by their exogenous versions. This is why, the following modifications of the initial reasoning were applied:

1. There is no measurement for SOX2 (because it either is not expressed in fibroblasts or the microarray failed to measure it). This is why, the endogenous SOX2 cannot be included in the optimization process and thus doesn't need any outgoing edges.

2. The measurements for POU5F1 are 0 except in the measurement where its own retroviral copy is present, i.e. in the POU5F1 transduction and in the 3TF and 4TF measurements. The effect of exogenous POU5F1 not being present (being 0) or endogenous POU5F1 being 0 are exactly the same which is why I don't need to consider the outgoing edges for POU5F1 either.

3. For KLF4 and c-MYC the endogenous versions have the same outgoing edges as their retroviral equivalents. In this way, whenever the retroviral transcripts of KLF4 or c-MYC are present, they overrule their endogenous versions. However, when they are not present but the endogenous genes show expression, they can still have a downstream effect.

After having applied all these modifications to the network, the only problem is that c-MYC now has 7 incoming edges, although, as mentioned above, only 6 can be handled by the software. Therefore, one more interaction had to be left out. In all the optimizations that had been run thus far, the interaction between IRS1 and c-MYC almost never had any importance in the optimization results, so it could be considered effectless and left out leading to the final network topology that is shown in Figure 4.2.

Now that the definite network structure is established, the new PKN can be loaded, the ensemble of all possible Boolean models can be created out of the PKN and the training of the model against the data can begin. For that purpose, an optimization function (see Equation 2.3) is used and minimized via a genetic (or evolutionary) algorithm. For details on genetic algorithms and their tweakable parameters such as population size, mutation rate, elitism, selective pressure, etc. please consult Fraser and Burnell (1970) and Crosby (1973) and Subsection 2.2.5.

Following the optimization the network will be continuously reduced by deleting the species that yielded poor fits and are thus likely to have a negative effect on the optimization. The fits with impaired scores may be due either to incompleteness or uncertainty of the literature data to build the network, or to outliers in the data points or even to the simplifications in biological complexity made in order to be able to carry out the optimizations.

However, the knowledge about which reduction steps improve the optimization contains a wealth of predictive power as to which interactions might be erroneous or need double-checking by experiments. The network together with the reduction steps made continuously is shown in figure 4.2.

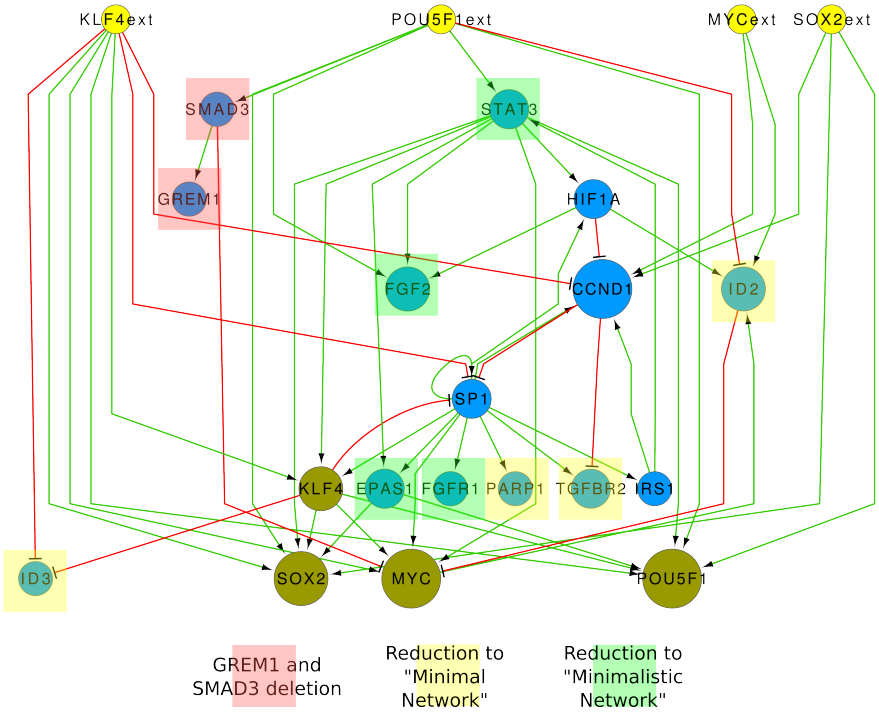

Figure 4.2: Whole Manually Reduced Pluripotency Network.

Edges: Transcriptional activation (Green), Transcriptional inhibition (Red)
Nodes: external vectors (Yellow). Pluripotency factors (Brown). Other transcription factors (Blue).
Also shown are the 3 continuous reduction steps carried out in the network. The transparent rectangles clarify which nodes have been eliminated to create the reduced networks whose optimization scores are shown in Table 4.3

Indeed, as briefly mentioned before, one will have to ask whether the genetic algorithm can even find a reasonable best solution in this vast optimization state space and what the goodness of the fit would be. Although there are no means to calculate the latter with CNO, there is still another very trustworthy measure for the question whether the algorithm finds the best solution. In fact, when sampling over random initial starting points for the optimization, one can easily compare the optimization results. I found that no matter how many optimizations I ran with random samplings over the starting point in the state space, the algorithm always found best models that were very similar to each other only differing in a few edges if any and always giving the same optimization score. This is a very strong argument

supporting the approach, because if the algorithm was not sufficiently accurate or the state space had many very different solutions which minimized the score function, the algorithm would find very different solutions and very different scores in such a vast state space.

In order to find out how similar the set of models inside the tolerance interval are, I compare them one by one to the best model using the similarity defined in Subsection 2.2.5 and then taking an algebraic average over all the fractions which leaves us with an average similarity of the models inside the tolerance interval towards the best model and thus also between each other.

Across all optimizations that will be carried out below for the different network versions, the similarity between models lying inside the tolerance interval of one optimization as well as the similarity of the best models between different optimizations of the same network is very high and never goes below 0.92. This means that either between two models inside the tolerance interval of one optimization or between two best models across different optimizations, at least 92% of interactions are identical. This is also another argument in favor of the applicability of the genetic algorithm: in fact, starting optimizations from different points in the optimization space always yields very similar results and the models inside the tolerance interval of one optimization are very similar to the best model that was found meaning that a very special set of models is found and the space of all models is strongly narrowed down.

4.3 Optimization of the Derived Model and Further Continuous Sensible Reduction of the Pluripotency Network

All optimizations that I will outline in the following have been carried out with the parameters presented in Table 4.2 and described in detail in Subsection 2.2.5. For every new network, I ran the optimization procedure 20 times, always starting at different random points in the optimization space in order to account for uncertainty in the process and reduce fitting errors by taking an average consensus model over all 20 optimizations as will be explained further below. In all the optimizations, the *Maximum Number of Stall Generations* always was the stopping factor as I intended it to be. I hypothesize that after 300 generations of unchanged best results, the actual best solution for the problem has been found.

It should be said in advance that a lot of what is known about networks in pluripotency will not hold for early reprogramming. This can already be seen when taking a close look at the data after 96 hours. This might

Table 4.2: Parameters for the Genetic Algorithm

The first parameter, namely the number of optimizations is not a parameter handed to the genetic algorithm but the real number of different optimization procedures run for each network. The 2 last parameters sometimes needed to stop an algorithm, when caught in a loop, were never used by the algorithm in my case, because it always found a solution ended by the maximum number of stall generations, i.e. the number of consecutive generations where the best result stays the exact same, i.e. the probability is high that the best solution has been found.

Parameter Name	**Value**
Number of Optimizations run per Network	20
Population Size	100
Probability of Mutation	0.7
Selective Pressure	1.2
Elitism	10
Relative Tolerance	0.05
Maximum Numbers of stall Generations	300
Maximum Time (s)	10,000
Maximum Number of Generations	100,000

be due to epigenetic mechanisms silencing specific genes at different stages of reprogramming (as will be further explained in Chapter 5), absence or presence of activating or inhibiting co-factors of transcription which might not be expressed yet but will only be expressed at later stages. Starting with the detailed description and interpretation of the complete manually reduced pluripotency network's optimization, I will analyze its continuous sensible reduction and the effects on the optimized models in the following. The different prior knowledge networks can all be deduced from the network representation and its reduction steps shown in Figure 4.2.

Optimization of the Complete Manually Reduced Pluripotency Network

The complete manually reduced pluripotency network as shown in Figure 4.2 reaches a best fitness score of 0.184. Considering that in some optimizations that I have run on the same data set with different curated networks, I had optimization score results as high as 0.40, this score already appears like a considerably improved value. However, when taking a closer look at the exact fit in Figure 4.3, one can see, that the fits still seem far from being perfect which should be explained in the following. It must be noticed that a minimum error of fitting is genuine and inherent for fits of continuous data to Boolean model outputs. More important in the analysis of such an ensemble of optimization graphs is the accordance of qualitative behaviors between model and data. Therefore, although very few fits appear to be

white corresponding to perfect fitting, there is much less that really are dark orange or red which correspond to erroneous behavior. Moreover, since the main focus of this Chapter lies on early reprogramming, I am especially interested in the *3TF* and *4TF* conditions which are the only combinations leading to reprogrammed cells, and these conditions appear to be fitted really well (last two rows of fits in Figure 4.3). While yellow fits, corresponding to errors below approximately 0.4, can sometimes still account for according behavior of model and data with just the inherent error of Boolean fitting to continuous data, in other cases of yellow fits, it is possible that the behavior of model and data are diverging. There are several reasons for this.

If we have a look at the fit for CCND1 in the *3TF* condition on the lower left of Figure 4.3 **A** for example, we can see that the score of the fit corresponds to a yellow color although the behavior is clearly different. While the model output suggests CCND1 to have low expression at the beginning which stays low after 96h, the data clearly suggests an up-regulation of CCND1. The fit is still shown in yellow because both start at approximately the same point so the fitting of time point 0 appears to be good. When now taking a look at the fit for KLF4 in the *SOX2* condition for example (Column of species KLF4 and second row corresponding to *SOX2* condition), the fit appears in a very similar yellow. In fact, in this case, the behavior is perfectly fitted. The expression has a certain value and more or less stays at this value. However, after rescaling, the value of the continuous data for KLF4 at this condition at time point 0 lie somewhere around 0.4. Since Boolean values can only take 0 or 1, there is already a rather big inherent error for the fitting at time point 0 which is complemented by another big inherent error at time point 96h. Therefore, the overall error of fitting is rather big although the qualitative behavior of the gene expression is perfectly fitted. When summing up across all species and conditions, these inherent errors, that in many cases might not reflect erroneous behavior, sum up to give an overall error reflected in the optimization score. For these reasons, one should not be influenced too much by the color code of the fits nor by the value of the score function of the fit because in order to find an optimal Boolean model, it is important that it can mimic the behavior inspired by the data. A perfect fit is not possible with a sensible normalization of the data.

In order to examine which edges exactly are the most important to reproduce the experimental data and thus might be the most important players in early reprogramming, I superposed the results for all the optimization runs that actually reach the same best score. As mentioned earlier in Subsection 2.2.5, each of the optimization procedures keeps track of the visited models and their fitness and at the end returns all models in the interval of a tolerance around the best score. When analyzing this set of models, the edges can be assigned a relative frequency of appearance that can be interpreted as a probability that the edge is needed to reproduce the experimental results.

The derivation of this probability is outlined in Subsection 2.2.5.

In the 20 superimposed optimizations, I filtered all the edges - that appear in the ensemble of optimized networks that have a score inside the interval of tolerance of the best model - for the ones that at least in one out of the optimizations have a probability of at least 0.9. In other words, in one of the optimizations, the edge in question needs to be present in 90% of the models that lie in the tolerance of the best model of that optimization. This filtering resulted in the optimized consensus model shown in Figure 4.2. The color code for nodes and edges are the same as in Figure 4.2 with the difference that there are now gray AND gate nodes resulting from the optimization. Moreover, the edges describing the interactions between the nodes now have different widths depending on the average probability of the edge to be present ranging from 0.11 to 1.0. These probabilities were calculated by taking the average of the edge probabilites described in Subsection 2.2.5 over all the optimizations reaching the same best score. When looking at the graphic, it quickly becomes clear that a few edges have a very high probability (some of them even being 1 across all optimizations) which accounts for a very high certainty that these edges are necessary in the model to be able to reproduce the data and can completely be relied on. The exact probabilities can be read from table A.2. It can be noticed in first place that due to the probability filtering, some of the species were deleted out of the final consensus model visible in Figure 4.3 **B**. These species are GREM1, ID3, PARP1 and TGFBR2. When taking a closer look at Figure 4.2, one can see that these species have very few incoming and no outgoing edges and thus have no downstream effects: they are themselves downstream targets of upstream regulators and their elimination out of the final interaction model does not strongly disrupt its connectivity. All the optimized interactions that will be thoroughly discussed in the following together with the fits of the species are shown in Figure 4.3. The highest confidence edges that need to be present to reproduce the data across nearly all optimizations have been found to be the following:

- SP1 (1) IRS1 (p=1.000)
- SP1 (1) HIF1A (p=1.000)
- SP1 (1) FGFR1 (p=1.000)
- SP1 (1) EPAS1 (p=1.000)
- HIF1A (1) FGF2 (p=1.000)
- POU5F1ext AND IRS1 (1) STAT3 (p=0.995)
- KLF4 (-1) SP1 (p=0.601)

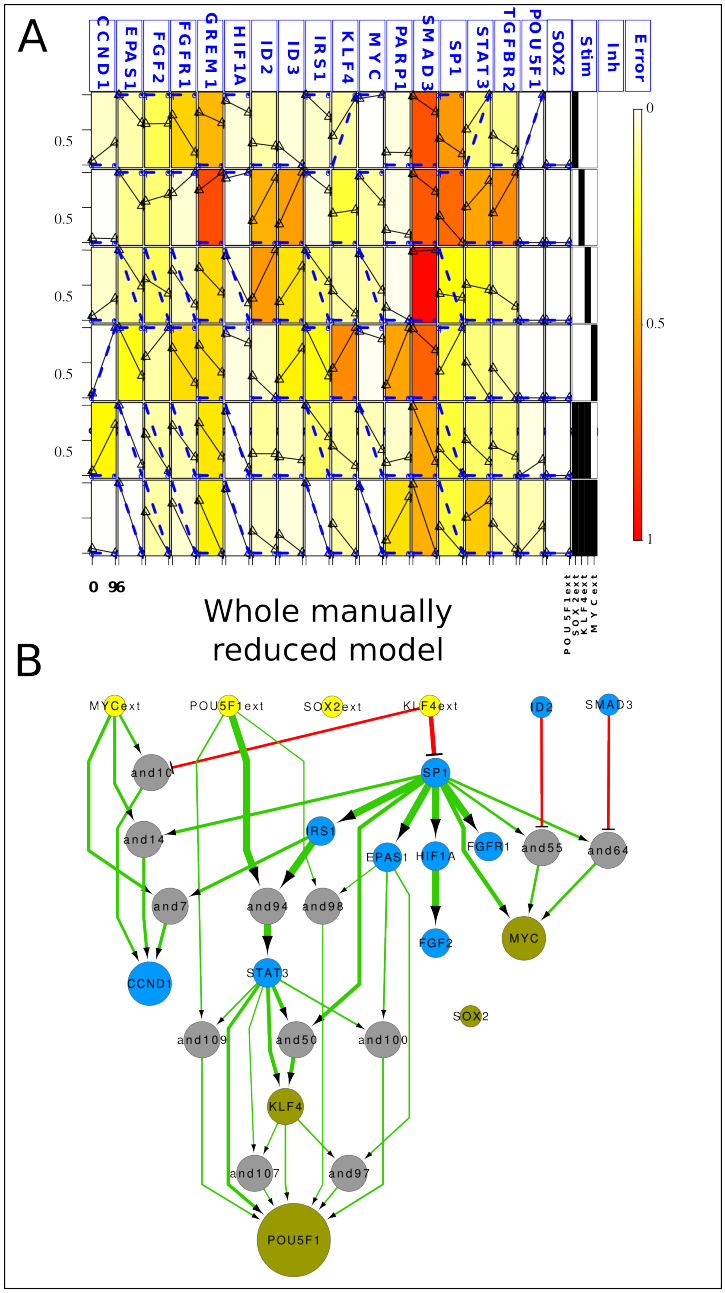

Figure 4.3: Optimized Whole Manually Reduced Pluripotency Network.

A Shown are all the fits of the single species for the different conditions. The blue dotted lines show the model output, while the solid black lines indicate the re-scaled data. **B** The optmization scaffold as described in the text shows edges with the highest probability of appearance in the optimization results (at least 0.9 in one of the optimizations). The edge line widths are mapped to the average probability of occurrence of the edge in all optimizations that reached the best scores. Edge and node color codes are the same as in Figure 4.2 with the difference that Boolean AND gates introduced via the optimization are marked in (Gray). The node size increases with the in-degree, i.e. the number of incoming edges to the node

No High Confidence Interactions Between Master Regulators of Pluripotency

To our surprise, none of the interactions that are known to be substantial in maintaining pluripotency, namely the self-sustaining interactions between the master regulators of pluripotency as postulated by Boyer et al. (2005), were found with a high certainty in the optimization of this model. On the one hand, this might be due to the fact, that the measurements for SOX2 have insufficient p-values (as already described earlier) due to lack of expression and that the p-values for POU5F1 measurements are not confident for some of the conditions which was why I set the measurements for it to 0 in these cases. On the other hand, this finding can be due as well to the early measurements at 96 hours after infection with the retroviral genes. The master regulators of pluripotency are essentially important for the maintenance of already pluripotent stem cells in their state, i.e. maintenance of self-renewal and pluripotency characteristics. In fact, it might be that in the first 96 hours of reprogramming, different other mechanisms and interactions play a more important role than the direct interactions between pluripotency master regulators, especially because the latter are still partially masked by epigenetic markers and thus transcriptional interactions will be unlikely to happen.

Interestingly, neither the retrovirally introduced *SOX2* gene (*SOX2ext*), nor the endogenous *SOX2* seem to play any role in our model in the first 96 hours of reprogramming as can be seen by the lack of incoming or outgoing edges into the nodes in the consensus model in Figure 4.3. However, since it has been shown, that unless SOX2 is already expressed in the cell lineage to be reprogrammed (e.g. neural progenitor cells, NPCs, in Eminli et al. (2008)), it is necessary for reprogramming, it is very likely to exert its full mechanism of action some time after the first 96 hours. As we know, the whole reprogramming process takes a longer time of at least 12 days (Takahashi and Yamanaka, 2006; Yu et al., 2007) depending on the cell lineage and reprogramming cocktail used and other experimental conditions and when time advances, as has been shown by Hanna et al. (2009), iPS cell colony number increases progressively. The two SOX2 versions are nonetheless included into the optimized consensus network (see Figure 4.3) in order to clarify their lack of importance in early reprogramming as found by my optimization.

A Pathway of Activation of Endogenous *POU5F1* and *KLF4* in Later Stages of Reprogramming

The externally driven impact with the highest probability is the activation of STAT3 by a combination of retroviral POU5F1 (POU5F1ext) together

with IRS1. This interplay hasn't been mentioned in literature before and might be interesting to be tested in future experiments. Basically, it hints to an interaction between the OCT4 transcription factor from the retroviral *POU5F1* gene (since the endogenous *POU5F1* is not expressed yet which is probably due to epigenetic masking or a lack of activating transcription factors) and IRS1 in order to induce *STAT3* expression which then in turn acts downstream on the endogenous *POU5F1* and *KLF4* expression. This could be one transcriptional mechanism of action to activate endogenous *POU5F1* in later stages of reprogramming. However, in early reprogramming (*3TF* and *4TF* conditions) at least in the rescaled data, there doesn't seem to be a strong activation of endogenous *POU5F1*. This is most likely due to the inhibition of *SP1* expression by KLF4 (which will be outlined further below). SP1 is in fact necessary to induce *IRS* as was found by the optimization with a probability of 1 (see enumeration above). However, when taking a closer look at the raw data in Section A.1, one can see that although the expression value for endogenous *POU5F1* is highest in the *OCT4* condition, there is a significant up-regulation of *POU5F1* in the *3TF* and *4TF* conditions as well. The strong discrepancy between the values, however, leads to the impression after rescaling that OCT4 stays at a low level in the reprogramming conditions and the algorithm still considers it as down-regulated due to the strong difference to the *OCT4* condition.

The results suggest that it is exactly the presence or absence of KLF4, IRS1 and STAT3 in an orchestrated manner that regulates this latter effect. In the *OCT4* condition, *IRS1* is expressed, leading to *STAT3* up-regulation via the above mentioned mechanism. Next, STAT3 can activate the endogenous *POU5F1* on its own or in combination with exogenous OCT4, endogenous KLF4 or EPAS1. At the same time, STAT3 can also activate downstream endogenous KLF4 on its own with a high probability of 0.490 or in combination with SP1 with an even higher probability of 0.519. KLF4 can then again activate downstream endogenous *POU5F1* directly or in combination with EPAS1. This strongly interconnected pathway enriched with positive feed-forward loops that starts from exogenous *POU5F1* and *SP1* could be a new pathway activating the pluripotency master regulators *POU5F1* and *KLF4* in later reprogramming stages.

However, as soon as KLF4 is present in the retroviral cocktail, i.e. in the *KLF4*, *3TF* and *4TF* conditions, KLF4 induced down-regulation of *SP1* leads to down-regulation of *IRS* which in turn leads to impaired activation of STAT3 and thus of downstream endogenous *POU5F1*. It is via this mechanism that the strongly counter-intuitive down-regulation of endogenous *KLF4* in the presence of retroviral KLF4 could be explained. The mechanism described here which was found exclusively by the optimization process, is unknown so far in its complexity and the relationship between species and although the optimization does not lead to perfect fits, the thus

found interactions are worth of being checked experimentally. It would furthermore hint to a negative effect of KLF4 in the reprogramming cocktail that can be hypothesized to strongly impair reprogramming efficiency by preventing earlier *POU5F1* activation. As will be discussed further below, KLF4 has other effects that are probably necessary for reprogramming. However, if it is experimentally possible to keep IRS1 or STAT3 expression high although SP1 is down-regulated, my optimized model suggests an improvement of the downstream *POU5F1* activation in early reprogramming and thus possibly a faster and more efficient reprogramming.

SP1: a Central Regulator in Early Reprogramming

As an interesting result, SP1 emerges to be a transcription factor with a strong importance in early reprogramming in my consensus model, i.e. most high probability edges have SP1 as a source or as a target. It is known, that SP1 plays an important and necessary role in early developmental embryos and has been associated with the maintenance of methylation-free CpG islands, cell cycle, and the formation of active chromatin structures (Marin et al., 1997). Moreover, the SP1/SP3 binding site which is present in a wealth of genes seems to be required for DNA demethylation (Simonsson and Gurdon, 2004), which is a necessary step in somatic cell reprogramming as will be further clarified in the probabilistic Boolean model we will develop in Chapter 5. In fact, during reprogramming an epigenetic reorganization of many genes will occur (Lister et al., 2009, 2011) which can also be deduced from the fact that iPSCs are in general transcriptionally more active than differentiated cells.

The finding that SP1 plays a crucial role is supported by the fact that in a model in which SP1 has been deleted, the optimization score deteriorates drastically to 0.316 (results not explicitly shown). Although SP1 itself does not appear to be perfectly fitted in the first two experimental conditions, it shows a slightly improved behavior in the *KLF4* and *c-MYC* conditions and its down-regulation in the model in the 2 last, the reprogramming conditions, is clearly reflected in the data. A thorough literature search for further upstream links of SP1 to improve its fits suggested an activation by STAT3 and by CCND1 (Tapias et al., 2008). New optimizations including these edges, however, left the score unimproved. The main focus, as mentioned before, lies on the analysis of the optimization results in the last two experimental conditions, which show very good fits throughout the data set and which are the essential reprogramming conditions.

It is important to notice, that SP1 is down-regulated by retroviral KLF4 with a relatively high probability of 0.601 (as mentioned above, the expectation value of probabilities is far below 0.5 which is why 0.601 is a high probabil-

ity value). Since the optimization considers the steady states of every model and since the main SP1 targets EPAS1, HIF1A, IRS1 and FGFR1 are all down-regulated in the *3TF* and *4TF* conditions, it is likely, that it is the down-regulation of SP1 which is the necessary step in early reprogramming. Contrary to the implications of SP1 in embryonic development described above, there might be an important feature in iPSCs that needs SP1 repression. In effect, it has been found that in most somatic cells, SP1 and SP3 work together to recruit histone deacetylase to inhibit human telomerase reverse transcriptase (hTERT) (Won et al., 2002). The activity of the latter is necessary for cells to acquire immortality and it is effectively expressed in ESCs and iPSCs (Rohani et al., 2013). Down-regulation of SP1 has been shown to induce activation of hTERT which is likely to be an important step in reprogramming. This might be one of the reasons why SP1 is down-regulated accompanied by down-regulation of FGF2, FGFR1, HIF1A and IRS1. However, the down-regulation of the 4 latter genes, should be thoroughly discussed because it is at least in parts a controversial result with up-to-date literature knowledge.

FGF2: High or Low?

In our data set, FGF2 shows little change of expression in the first two (*OCT4* and *SOX2*) conditions, a slight down-regulation reflected in the model in the *KLF4* condition, a slight up-regulation in the *MYC* condition and a pronounced down-regulation reflected in the model output in the two reprogramming (*3TF* and *4TF*) conditions which are the conditions of interest for the following analysis.

FGF2 has been shown to be a crucial mitogen whose expression is necessary for hESC self-renewal and to prevent differentiation (Greber et al., 2007b,a). This finding seems to contradict our data and the optimization result at first view because FGF2 is down-regulated in the reprogramming conditions which eventually lead to iPSCs. However, it can be controversially discussed: As was shown in the introductory Subsection 1.2.2 in Figure 1.4, FGF2 can maintain self-renewal of hPSCs at low as well as at high concentrations via a mechanism including PI3K and MAPK/ERK signaling (Dalton, 2013). The PI3K signaling pathway lies downstream of IRS1 which is an activator of the pathway and is included in my consensus model as well. At low expression levels, FGF2 activates self-renewal via MAPK/ERK signaling. At high levels it also activates PI3K signaling which down-regulates MAPK/ERK via AKT thus keeping MAPK/ERK levels in a tight range favorable for self-renewal. As mentioned above, IRS1 can activate PI3K as well. High levels of IRS1 combined with low levels of FGF2 might thus down-regulate MAPK/ERK in a strong manner. It is assumed that low levels of ERK signaling support

stem cell maintenance while elevated levels stimulate differentiation (Dalton, 2013). This theory is supported by our findings that FGF2, FGFR1 and IRS1 are down-regulated in the 2 reprogramming conditions. Although it has been found that IRS1 is down-regulated during mESC differentiation and that its down-regulation is associated with a decrease in Oct4 levels, it is well known that there are substantial differences in mouse and human ESCs especially when it comes to the signaling pathways (Schnerch et al., 2010). Thus, the IRS1 down-regulation by KLF4 in our model might be orchestrated with the FGF2 pathway described above (see Figure 1.4): it is possible that the down-regulation of IRS1 counteracts the down-regulation of FGF2 in the MAPK/ERK regulation. High levels of IRS1 would lead to a too strong down-regulation of MAPK/ERK possibly impairing self-renewal.

Hypoxia Inducible Factors: an Attempt to Reconcile Gene Expression and Protein Regulation

The next controversy to be discussed is the down-regulation of HIF1A and EPAS1, two sensors of hypoxia that are known to play an important role in the embryo residing in 3-5% oxygen conditions (Forristal et al., 2010). When exposed to hypoxic conditions, cells need to up-regulate parameters of oxygen-dependent reactions to ensure sufficient levels of species involved in them. This is where hypoxia inducible factors (HIFs) come into play which are master regulators of over 200 downstream target genes involved in erythropoiesis, apoptosis and proliferation (Semenza, 2000). It has furthermore been shown that hypoxic conditions enhance reprogramming of murine ESCs (Yoshida et al., 2009). However at the same time it was found that hypoxic conditions alone are able to induce the differentiation of human ESCs into functional endothelial cells (Prado-Lopez et al., 2010) in one publication, and in another that low oxygen tensions prevent differentiation of hESCs (Ezashi et al., 2005). In brief, it can be summarized that advantages of culture in hypoxia are controversial and literature findings in different high impact journals seem to be contradictory at first sight (Chen et al., 2009; Forristal et al., 2010). I will try another interpretation in the light of my new findings concerning the regulation of HIF1A and EPAS1.

HIF1A and EPAS1 are regulated at the protein level in an oxygen dependent manner: Under atmospheric 20% oxygen conditions they are hydroxylated by prolyl hydroxylases (PHDs) and subsequently degraded by the von Hippel Lindau complex (VHL) (Semenza, 2003). At hypoxic conditions, PHDs are unable to hydroxylate HIF1A and EPAS1 and their protein levels increase. Our experiments have all been carried out at atmospheric oxygen levels (normoxia). It is thus surprising at first, that HIF1A and EPAS1 both seem to be expressed as well in fibroblasts as in fibroblasts transduced with GFP

(HIF1A in a stronger manner, EPAS1 less strong as in FIB as can be seen in Table A.1). However, since they both have to be able to react quickly to oxygen changes in the environment and are regulated at the protein level, it could be hypothesized that there is always a pool of mRNA (which is the quantity that is measured in our microarray expression profiles) present and steadily transcribed which is then translated to protein and degraded directly, thus leading to a dynamic equilibrium. When the degradation stops in hypoxic conditions, protein levels quickly accumulate to deliver a strong and fast response. Now, why are both genes significantly down-regulated as soon as KLF4 is present in the reprogramming cocktail?

It appears that it is again the down-regulation of SP1 that plays a fundamental role. While it was an important factor keeping transcription levels of HIF1A and EPAS1 high, it is down-regulated by retrovirally introduced KLF4 protein which induces subsequent down-regulation of HIF1A and EPAS1. Since culture was carried out in normoxia, there is no need for an up-regulation of HIF1A and EPAS1 when cells are transformed into iPSCs. However, it is possible that due to this down-regulation of the hypoxia inducible factors, cells will have a slower response to hypoxia at this stage. As it is known that hypoxia enhances reprogramming (Yoshida et al., 2009), this mechanism is likely to take place at a later stage of reprogramming than in the first 96 hours because in this period of time the hypoxia inducible factors are strongly down-regulated and less likely to be responsive to hypoxia. Another possibility is that via more complex mechanisms, a down-regulation of the hypoxia inducible factors would be prevented in hypoxic conditions.

The result of my optimization placing a highly confident activation of FGF2 by HIF1A gives reason to believe that a later reprogramming mechanism could work through the positive FGF2 effect on hESC and iPSC culture (Greber et al., 2007b,a). Since ESCs and iPSCs are highly responsive to hypoxia as well (Prado-Lopez et al., 2010; Ezashi et al., 2005), it is thus reasonable to hypothesize that there is an intermediate developmental state in which the hypoxia inducible factors are down-regulated while they are up-regulated in the pluripotent as well as in the fibroblast states. It seems to be necessary for the cell to pass by this intermediate state during reprogramming and it is possible that it does so as well during differentation. As a last thought, it should be mentioned that in the publication Prado-Lopez et al. (2010) claiming hypoxia induced differentiation of hESCs, an FGF2 free medium was used (at least the method section only mentions culture on human foreskin fibroblasts (HFF) extracellular matrix (ECM) which does not contain FGF2 and needs addition of 100 ng/mL in order to keep hESCs in culture as stated in Meng et al. (2010)). Therefore, it is possible that the observed differentiation is not hypoxia induced but is due to a lack of FGF2 which would support my hypothesis.

After having thoroughly discussed the latter controversial results of the high probability interactions found by the optimization, I will now turn to the less probable interactions that have been found. Although the other interactions not mentioned in the enumeration above have a lower average probability across all carried out optimizations, they still have a probability above 0.9 for at least one of the optimizations because that was the criterion of selection for the edges. In other words, there exists at least one set of models from an optimization in which the edge in question has a very high probability of appearance or in other words a very high confidence of being necessary to reproduce the data. Moreover, due to combinatorial and size regulation facts mentioned above which strongly lower the expectation value of the relative frequency of an edge, an average of above 0.1 still seems elevated in comparison to the majority of other logic gates which are way below this value. Therefore, the other regulations should be taken into consideration as well, such as the one for CCND1 which is one of the periodically changing factors of the cell cycle. Interestingly, this gene shows a strong basal expression in the raw microarray data. However, since in most conditions it will be strongly up-regulated and only shows a slightly lower expression in the *4TF* condition, the rescaling normalization resulted in the time point 0 measurement being close to 0. It should be mentioned at this point again, that the fits should not be read as absolute expression but as relative expression changes in early reprogramming.

Keeping the focus on hypoxia inducible factors, it can be seen from Figure 4.3 that EPAS1 is multiply involved in the activation of endogenous *POU5F1* together with different other factors. EPAS1 works in cooperation with STAT3, KLF4 or exogenous POU5F1 with the respective probabilities of 0.189, 0.151 and 0.108. It cannot be known whether these probabilities are mutually exclusive or can be added up to form a total probability of involvement for EPAS1 in the induction of *POU5F1*. However, these interactions on their own already constitute a good indicator for a possible downstream action of EPAS1 on *POU5F1*. It is a surprising result though that despite culture in normoxia that abolishes induction of HIF1A and EPAS1, the optimization finds the activation of *POU5F1* by EPAS1 to be of importance. Since EPAS1 and HIF1A are strongly expressed in fibroblasts at atmospheric O_2 conditions, I tried an explanation for this phenomenon including the protein dynamics of EPAS1 and HIF1A further above. However, due to a lack of protein data, I can only hypothesize this flux balance hypothesis to be true which would favor the assumption that EPAS1 cannot activate POU5F1 under normoxia because its protein levels are too low. The rather low probabilities in comparison to the high confidence interactions would favor this hypothesis. Nonetheless, if the regulation on the protein level is not absolute, i.e. would not completely abolish EPAS1 downstream action, it is possible as well that the interaction plays the slight role suggested in the optimization.

Reasons for Discarding SMAD3 and ID2 and Evidence for the Regulation of Endogenous c-MYC by SP1

Beside its already mentioned powerful involvement in controling many downstream targets involved in reprogramming, SP1 appears to exert another role in the regulation of the endogenous *c-MYC* gene. The optimization found SP1 to be able to activate *c-MYC* on its own with a probabilty of 0.568 while there is evidence for 2 other mechanisms including repression of *c-MYC* by SMAD3 in conjunction with SP1 activation of *c-MYC* and repression of *c-MYC* by ID2 in conjunction with SP1 activation of *c-MYC* with the respective probabilities of 0.314 and 0.294 (see Table A.2). The first predominant direct interaction is likely to be responsible for the down-regulation of *c-MYC* in the presence of KLF4 in the retroviral cocktail through the down-regulation of SP1 which is very well reflected in model output and data. It can be hypothesized that the bad fits for SMAD3 that reflect a strong discrepancy between model output and real data situation are responsible for the inclusion of the SMAD3 action on *c-MYC* which will be the main reason for the elimination of SMAD3 to reduce the model in the following Subsections. Therefore, I will not discuss this interaction here. Although ID2 seems to be fitted in a slightly better way, in reality the information about the upstream regulation of ID2 seems to be insufficient which is why the model output considers it as constantly down-regulated, i.e. unregulated in the model. Since this seems to be a good fit for the optimization, it unfortunately includes the interaction into the model. Therefore, just as SMAD3, it will be eliminated in the reduction step of the model and not further be discussed here. The only interaction to be discussed is therefore the direct one of SP1 activating *c-MYC*.

This interaction has been known for a long time (Geltinger et al., 1996; Majello et al., 1995) but to my knowledge has not been further analyzed with respect to reprogramming and its importance in iPSCs or hESCs. It has been mentioned before that SP1 and SP3 represses hTERT transcription via recruitment of HDACs (Won et al., 2002). However, there is also evidence for the cooperation of SP1 with c-MYC to activate downstream hTERT which is thought to be a crucial step in the acquisition of immortality of stem cells (Kyo et al., 2000). In the light of the new findings from the optimization it appears that early reprogramming down-regulation of SP1 is necessary which we hypothesized amongst others to be related with its release of the hTERT promoter and thus epigenetic demasking in order for hTERT to be prone to transcription. The cooperation of SP1 with c-MYC to activate hTERT transcription could hint to a more complex regulatory mechanism: As long as SP1 is down-regulated in early reprogramming, hTERT is epigenetically unmasked but possibly not transcribed due to lack of activating transcription factors. Once SP1 is up-regulated again which will probably occur in

the later stages of reprogramming, c-MYC could be up-regulated directly by SP1. Together, SP1 and c-MYC could then induce hTERT transcription. It must be asked, however, why hTERT will not simply be epigenetically masked again via binding of SP1 and recruitment of HDAC. I will offer a few possible explanations for this: First, it is known, that SP1 together with other SP1 proteins can form homomultimeres which activate transcription synergistically (Pascal and Tjian, 1991) but may not be able to bind HDAC to silence hTERT again. Another possible explanation involves differentially expressed transcriptional or epigenetic co-factors. It could thus be, that SP1 and SP3 are needed for the binding of HDAC and that SP3 will be down-regulated in later stages of reprogramming. As a last and interesting possibility, it could be that HDAC is inhibited in later stages of reprogramming. The fact that valproic acid (VPA), a HDAC inhibitor, is known to enhance somatic cell reprogramming 100-fold (Huangfu et al., 2008), strongly supports this hypothesis. In this case, SP1 would again emerge as a very important factor whose dynamic expression during reprogramming needs to be tightly regulated. A first down-regulation in order to release the epigenetic marks at the hTERT promoter would thus be followed by another up-regulation in order to activate hTERT and the other above mentioned downstream target genes transcription.

A Complex Activation/Repression Pathway of CCND1 via Retroviral MYC and KLF4 and Endogenous SP1 and IRS1

Another interesting feature found by the optimization is the complex regulation of CCND1, a cyclin involved in cell cycle regulation, by retroviral KLF4 and c-MYC involving SP1 and IRS1 as mediators. In the rescaled data compared to the model outputs in Figure 4.3, it appears that in the first 3 conditions, CCND1 doesn't change its expression strongly, then gets strongly up-regulated in the *MYC* and *3TF* conditions, the first one being in accord with the model output, the second in disaccord with it and in the *4TF* condition it again stays at low expression levels. It should be noticed however, that when taking a closer look at the data in Section A.1, CCND1 has by far the highest expression rate of the measured genes in all assays. Therefore, although there are relative changes of expression showing an increase when KLF4 is absent and c-MYC present, the expression is still high as well in the presence of KLF4. It can only be remarked that in the presence of KLF4, the expression is not significantly different from the one of the GFP-transduced fibroblasts control assay.

It has been found earlier that KLF4 represses *CCND1* expression through the SP1 binding motif via competition with SP1 proteins on the *CCND1* promoter (Shie et al., 2000). This mechanism is reflected in our optimized

model via the complex interactions between retroviral KLF4, SP1 and exogenous c-MYC. Retroviral KLF4 represses *CCND1* in an exogenous c-MYC dependent manner and at the same time inhibits *SP1* expression which has the potential to activate *CCND1* in a c-MYC dependent manner as well. As soon as KLF4 is present, *CCND1* will be repressed. In the absence of KLF4, CCND1 can be activated via the interaction of retroviral c-MYC with SP1 or IRS1.

As a summary, retroviral KLF4 action on SP1, IRS1 and CCND1 appears to be the main mechanism that leaves *CCND1* expression unchanged in comparison to GFP-transduced fibroblasts. Since *CCND1* repression has been related to inhibited cell proliferation with a G_0/G_1 arrest (Shimizu et al., 2010), it could be hypothesized that early reprogramming requires this cell cycle arrest in order to establish new transcriptional and epigenetic profiles before mitosis partly erases these latter marks. Another reason for this arrest could be to prevent uncontrolled proliferation as cancer and stem cells have similar features. However, it should be taken into consideration, that although *CCND1* expression is unchanged in the *4TF* condition, this is in comparison to GFP-transduced fibroblasts which already show a strong expression of *CCND1*.

TGFBR2, PARP1, ID3, GREM1: Why Don't They Play a Role?

Finally, there are the above mentioned species that have been left out of the final consensus network due to the low probability of the interactions they are involved in. Due to the lack of interactions, these species are regarded as down-regulated at all time by the optimization. Although there is no consensual information about these species, it is still possible to analyze the time course of the data of these species and try to find a sensible explanation for it without the claim of a systems biology justified hypothesis. Starting with TGFBR2, one can see from the fits in Figure 4.3 **A**, that it is down-regulated in all conditions except for the *SOX2* condition with the strongest decrease taking place in the *MYC* and *4TF* conditions. It has been found that reprogramming requires a mesenchymal-epithelial transition (MET) (Samavarchi-Tehrani et al., 2010) which is inhibited by the TGFβ pathway that favors an epithelial-mesenchymal transition (EMT) (Li et al., 2010). A down-regulation of TGFBR would lead to a decrease in TGFβ signaling which favors the MET necessary for reprogramming. It is interesting to notice that the presence of c-MYC enhances the down-regulation of TGFBR which could be one reason why this experimental condition is more efficient than the *3TF* condition. To my knowledge, no relationship between TGFBR and c-MYC has been described in literature thus far.

As described earlier, since PARP1 is mainly an epigenetic regulator, I only

included upstream regulation of PARP1 in order to mainly keep the regulatory network based on transcriptional interactions. This resulted in the only remaining interaction including PARP1 to be the activation by SP1. PARP1 being strongly activated only when c-MYC is present in the reprogramming cocktail, there seems to be an unknown interaction involving c-MYC that leads to PARP1 activation in the one or the other manner. Due to lack of knowledge about upstream regulation of PARP1 and the need to keep the model simple and based on transcriptional interaction, PARP1 will be eliminated out of the model in the reduction steps below.

Although ID3 was found to be differentially regulated, the only curated interaction that could be found was the inhibition of ID3 by KLF4 (Nickenig et al., 2002). After optimization, this interactions and thus ID3 is eliminated out of the model. It has been found that ID2 and ID3 need to be down-regulated in order for cells to induce an EMT (Valcourt et al., 2005). Since it has been shown, that cells undergo a MET in fibroblast reprogramming, it would be reasonable to assume that ID2 and ID3 need to be up-regulated in early reprogramming. However, neither the *3TF* nor the *4TF* combinations show an up-regulation of the 2 but on the contrary, rather a down-regulation. Moreover, the diverging dynamics of the other assays - up-regulation of both ID2 and ID3 in the *SOX2* and *KLF4* conditions which strongly contradicts the postulated inhibition of ID3 by KLF4 (Nickenig et al., 2002) - demonstrate that there are missing upstream regulatory mechanisms for ID2 and ID3. This could be due to the lack of included TGFβ signaling. In fact, *ID2* and *ID3* are both downstream target genes of this pathway and are repressed by the TGFβ branch and induced by the BMP4 branch of this pathway (Kowanetz et al., 2004). Since it is not possible to include protein signaling pathways into the model in a sensible manner, it is thus reasonable to exclude the IDs from the model in further reduction steps.

Without discussing GREM1 in greater detail, SMAD3 deletion out of the model will induce GREM1 deletion, because GREM1 is only linked to SMAD3 in the network. It should finally be mentioned that all the interactions that were present in the PKN in Figure 4.2 and that are not found to be substantial in the final consensus network in Figure 4.3 B, can be regarded as not necessary to reproduce the experimental data with a very high certainty. In other words, it is highly unlikely from our data set and literature network, that these lost interactions play a role in early reprogramming which possibly gives reason for some of the interactions to be double-checked experimentally.

Elimination of Non-Confident Interactions and Species Leads to a Minimal Network of Early Reprogramming

Two of the species that are noticeably badly fitted in Figure 4.3 are GREM1 and SMAD3 under all conditions. In an attempt to progressively construct a confident minimal interaction network of early reprogramming and pluripotency which could be tested in time course simulation experiments eventually, I eliminated the latter species in the next step to examine the effects on the optimization process. Since GREM1 has no downstream targets at all and SMAD3 is only involved in downstream inhibition of MYC which has many other inputs as well, their deletion might not influence the fitting of downstream targets negatively and is likely to improve the score of the optimization significantly.

In effect, deletion of GREM1 improves the optimization score by approximately 0.014 to 0.170 for the *GFP* measurements. As expected, since GREM1 doesn't affect any downstream targets, the remaining fits and the results stay the same as before (results not shown). Since the overall fitting for SMAD3 is apparently the worst in the optimizations so far, SMAD3 has been erased from the model as well. This time, the deletion improves the optimization by approximately 0.027 to 0.137. This stronger improvement than with the GREM1 deletion is due to the worse fitting of SMAD3 as can be seen in Figure 4.3. While for GREM1 only the *SOX2* measurement is red showing incoherent behavior of model and data, SMAD3 is badly fitted for nearly all measurements, especially for the first 4 measurements with the single factors. In fact, as described in literature, SMAD proteins act on transcription after having been activated via phosphorylation. Hence, it is not only the amount of SMAD proteins that is important for the transcriptional activation but the amount of phosphorylated SMAD proteins. Since this entity has not been measured, but only the expression of the SMAD3 gene or more precisely the quantity of gene transcript, it is justifiable to leave SMAD3 out of our network completely (Frederick et al., 2004). In addition, in order to build a minimal model for early reprogramming with factors playing an important role, the 4 above mentioned genes have been left out, that don't have interactions with high confident probabilities.

In order to further improve the optimization and build a confident interaction network that might be able to describe early reprogramming, I further reduced the network, leaving out all the above mentioned species that still have measurements with errors around 0.5 and that have little downstream influence, including both ID proteins ID2 and ID3, PARP1 and TGFBR2 besides GREM1 and SMAD3. PARP1 is in fact an epigenetic modifier and might work at a different time scale than transcriptional mechanisms, TGFBR2 has no downstream influence since it is not a transcription factor, but is

only transcriptionally controlled. The same is valid for ID3. Since all other species stay equally well fitted after the deletion (see Figure 4.4), the ID2 downstream effect cannot have been strong. Leaving out 6 species in total the score now improves from 0.184 to 0.118 (see Table 4.3).

As can be seen when comparing Figures 4.3 and 4.4, the optimized networks are very similar and nearly all the edges found to be present in the consensus network are the same. Due to the reduction of the network, a few of the AND gates now have different names due to their reduced number but when the interactions are still the same, the AND gates in question are as well the same as before. In general, it can be noted that probabilities are slightly shifted (see Table A.3).

Naturally the interactions of species that have been left out, i.e. ID2 and SMAD3, are not present anymore. However, as could have been expected, their influence on downstream c-MYC is fully compensated by its direct activation through SP1 which now has a probability of 1 (see Table A.3) compared to 0.568 before. The fact that this interaction fully compensates the other 2 that were found in the previous consensus network is confirmed by the same fitting for c-MYC in both optimizations.

Another exception is the new interaction between retroviral OCT4 and endogenous KLF4 to induce endogenous *POU5F1* expression. However, this interaction only has the lowest probability of all included edges of 0.127. This is probably due to the shift of probabilities which has occurred because many species have been eliminated and thus also their probabilities (which, although very low on their own could probably make up a non-negligible value when added up and re-distributed on fewer interactions). These probabilities will be re-distributed among the remaining interactions. Together with the finding, that there is only one new interaction, this is why a lot of the interactions have now higher probabilities. This is, however, a finding that strongly supports the optimized model and its reduction because it means that the filtered high probability edges are not replaced with new interactions but stay in the model and the score improves by attributing them the probabilities from the erased badly fitted species and interactions.

Minimalistic Pluripotency Network With SP1

I further reduced the model in order to break it down into a few crucial interactions between the main master regulators of pluripotency and to broaden our understanding of the first steps that have to happen in early reprogramming. It is interesting to see that the crucial interactions don't have a lot in common with the thus far accepted networks of Boyer et al. (2005) and others which is a sensible result since these interactions have been found

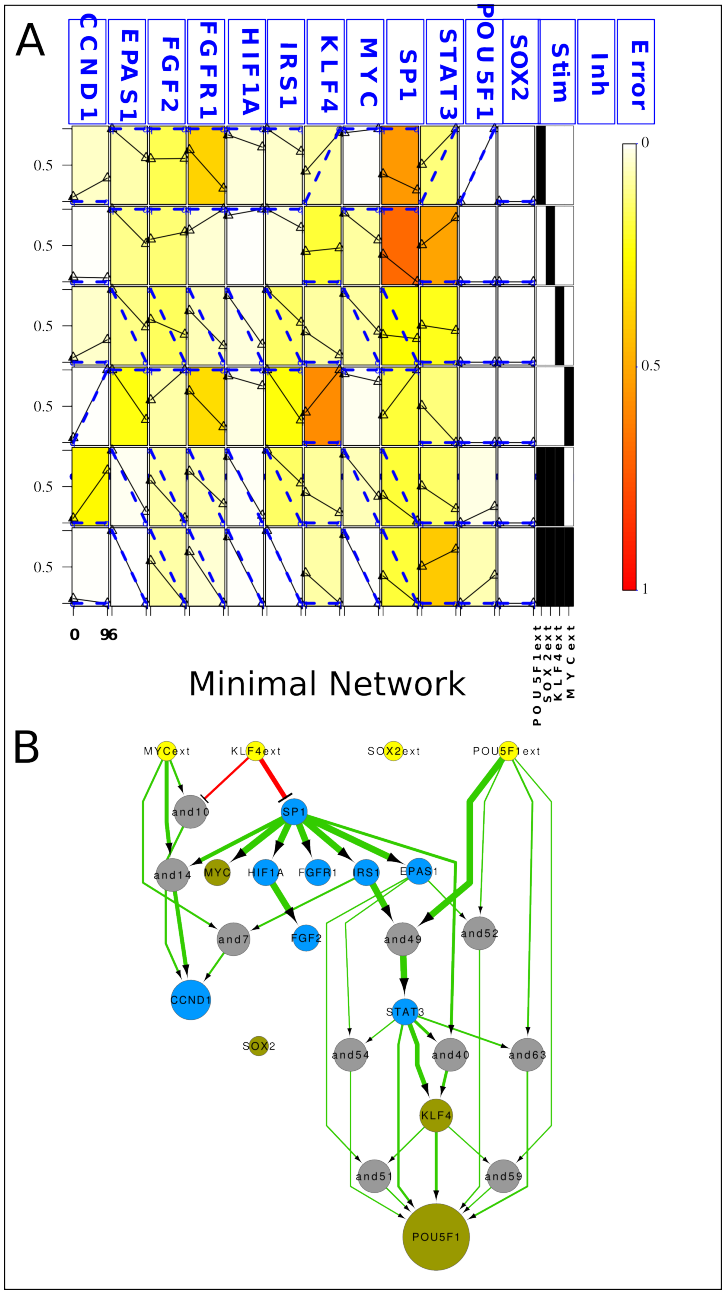

Figure 4.4: Optimized minimal Network.

The network has further been reduced by ID2, ID3, PARP1 and TGFBR2 for reasons described in the text. All color and size codes are the same as in Figure 4.3

to be important in already pluripotent cells. In principle, the minimalistic model shares a lot of features with the minimal and the big model. The double inhibition by KLF4 is still present, two activations of CCND1 by ex-

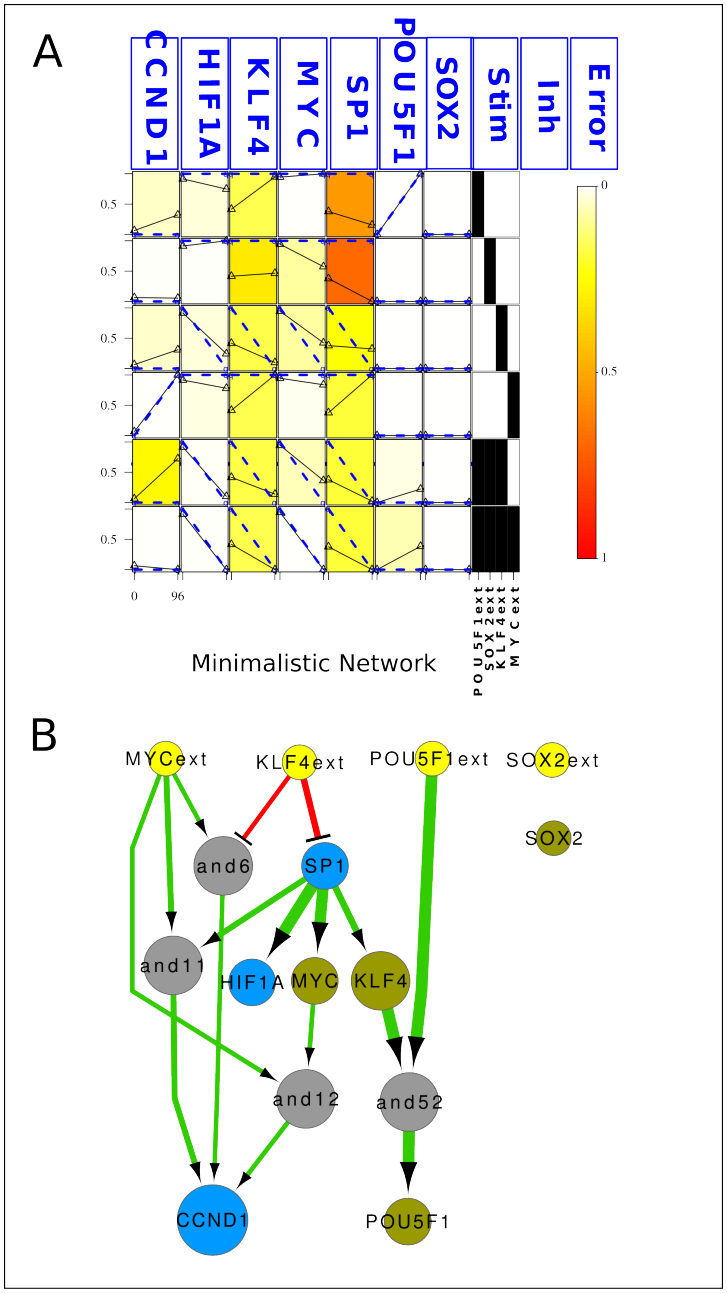

Figure 4.5: Optimized minimalistic Network.

The network has further been reduced by EPAS1, FGF2, FGFR1, IRS1 and STAT3 for reasons described in the text. All color and size codes are the same as in Figure 4.3

ogenous MYC and the downstream activations by SP1 show up. However, there are also new interactions: endogenous KLF4 together with retroviral OCT4 can activate endogenous *POU5F1* with the highest probability of

1.000 (see Table A.4). Since the species are not worse fitted than in the complete model and the optimization score strongly improves, it appears that in the minimalistic model KLF4 on its own is able to take the place of the complex activation pathway of endogenous *POU5F1* via STAT3, IRS1, EPAS1 and KLF4 that I have explained in detail above. Another new interaction is induced by the re-inclusion of the interaction between c-MYC and CCND1 (which had to be left out in the other networks due to the incoming edges constraint for CCND1) and revealed a non-negligible importance for this edge. This interaction could possibly have been compensated for in the optimizations of the bigger networks by the exogenous *c-MYC* gene. However, it now appears that retroviral as well as endogenous c-MYC cooperate in the activation of *CCND1*. It should be noted that CCND1 has a strong periodicity throughout the cell cycle. Since a certain distribution of CCND1 in a population thus represents a certain distribution of cell cycle stages, I have to assume the same distributions across the different samples in order to compare them. If we assume the same distributions, averaging across the samples gives us representative results for CCND1. It is still possible, however, that due to its periodicity CCND1 underlies stronger fluctuations in its expression value. Moreover, due to the Boolean optimization approach, it is sometimes not possible to unravel which out of several mechanisms is the crucial one. In order to discriminate between them, one would have to carry out further experiments.

For the first time, the auto-regulatory loops of the endogenous species SP1, KLF4, SOX2, POU5F1 (see Table 4.1) were tested in this model as well. As mentioned above, they could not have been included into earlier networks. However, since they are not part of the consensus network, they don't seem to play a role in early reprogramming in this minimalistic network. It is very unlikely that they play a role in the bigger networks where there is even more interactions to compensate for them. It is an interesting result to unravel that in addition to the lack of most mutual activations between endogenous master regulators of pluripotency postulated by Boyer et al. (2005), neither the self-sustaining loops of the transcription factors seem to play a role in the first 96 hours.

The optimization results for the different model variants are summarized in Table 4.3.

Table 4.3: Optimization scores for the different model variants, each for normalization against *FIB* and *GFP*

Model variant	***GFP* score**	***FIB* score**
Whole Manually Reduced Network	0.184	0.192
Deletion of GREM1 and SMAD3	0.137	0.145
Minimal network	0.118	0.132
Minimalistic network	0.098	0.114

4.4 Simulation of the Optimized Network in a New Boolean Network Simulator

The following section is partly based on our publication: Bock, Scharp, Talnikar, and Klipp (2013)

4.4.1 Presentation of BooleSim: An in-Browser Boolean Simulation Tool

After having built a bigger transcriptional interaction network and optimized a Boolean model of this network using early reprogramming microarray data, another necessary step in the work of a theoretical biophysicist is the simulation of the model. As part of the qualitative analysis of a network's dynamical behavior in the first approaches to a biophysical problem or in the last part when it comes to model testing and predictions, time course simulations have been and always will be a crucial step in the framework of Systems Biology.

When it comes to Boolean networks, there are a few software tools available for download that include a wealth of functionality and are very suitable for more extended use, i.e. the booleannet package for python (Albert et al., 2008) or the BoolNet package for R (Müssel et al., 2010). However, both tools require a minimum knowledge of the programming languages they come with, i.e. python or R and also some scripting abilities. Thus, when having a network model ready, a quick analysis of the network behavior in different situations, i.e. time course simulations from different initial conditions or different network versions, are not easily feasible. Therefore, we wrote BooleSim, a cross-platform, in-browser Boolean network simulator (Bock, Scharp, Talnikar, and Klipp, 2013). BooleSim is an open-source tool, that allows import of Boolean networks in different common file formats. It is easily accessible and enables in-browser network visualization and time course

simulation, on-click manipulation and export of graphical views, time series and network files.

The import and export file formats mentioned above include the boolean-net formalism, the BoolNet formalism and the more recent jSBGN format (Krause et al., 2013). It is moreover possible to create a new network from scratch inside the tool and input the Boolean update rules in the *Rules Tab* (see Figure 4.6). All the input formats are text-based and consist of the Boolean update rules defining nodes and edges. When importing a model, the tool creates a node for every variable in the model file. In the following, edges are created between every two nodes depending on the Boolean functions defining which nodes influence each other. A positive interaction corresponding to stimulation or activation in biological systems is represented by an arrow, a negative one corresponding to inhibition is represented by a *T-shaped* arrow.

BooleSim uses the modern web technologies HTML5 and JavaScript in order to make the Boolean network simulator as interactive as possible. The network graph is rendered as a dynamic SVG, generated during runtime using the biographer-UI (Krause et al., 2013). The network layouting is based on the d3 gravity/repulsion algorithm (http://d3js.org/) and arranges the nodes in a force-directed manner depending on the size of the labels and the length of the edges.

Inside the network graph, the color of each node represents its binary state. Yellow corresponds to active (1) and blue to inactive (0) states. Left-clicking on a node, changes its state, right-clicking node deletes it, while right-clicking on the canvas outside a node creates a new node. Boolean update rules can be seen when hovering over the node in question and can be edited in a separate text box tab. Manipulation of the update rules is internally evaluated, translated into a new network layout and applied in the algorithm in the next simulation steps.

During a simulation, nodes change color according to their changing states. This transition is happening in a a smooth transition effect based on jQuery features (http://jquery.com/). The simulation terminates as soon as a steady state is reached. If the system is trapped in a cyclic attractor inducing oscillations, the simulation continues as long as the user doesn't stop it via clicking the Simulate/Pause button.

The progression of the node's state is dynamically represented in the time series tab (see Figure 4.7): Node names are placed in one column on the left side of the time course, their time dependent states are shown in yellow or blue in the heatmap representation with time progressing with the green arrow (green in the software tool, black in Figure 4.7 to the right. This heatmap time course can also serve to detect cyclic or point attractors.

As mentioned above, BooleSim also supports exporting networks to Python BooleanNet or R BoolNet text file formats, as well as export to and re-import from biographer's exchange format jSBGN. This latter format allows for storage of the graph alongside its update rules. Moreover, network graph and time series can also be exported as an SVG vector image file.

4.4.2 Simulation of the Optimized Minimalistic Pluripotency Model Using BooleSim

In order to demonstrate the functionality and usefulness of our tool and still keep the simplicity of a small model, I'm going to use BooleSim to analyze the minimalistic Boolean network that I derived from the optimization process in section 4.3.

First of all, the Boolean network needs to be translated into one of the importable file formats that BooleSim can handle, i.e. the booleannet, BoolNet or jSBGN text files or to be handed to the tool via the aforementioned *Rules Tab* in the browser. For simplicity, I implemented the model of the optimized minimalistic consensus network directly in the browser in the JavaScript syntax. The JavaScript syntax for the implementation of Boolean networks is shown in the tab where the user can input the Boolean update rules. The implementation of the rules for the *4TF* condition is shown in Figure 4.6. It should be noted that it is not the one best model found by the optimization algorithm that is implemented here but the consensus model that was reached through complex filtering of all the interactions included in the interval of tolerance around the best model (for thorough description, refer to section 4.2).

The Boolean models corresponding to the different external conditions were then simulated in BooleSim until they reached a steady state. The time courses are shown in Figure 4.7. When comparing them to the model fits for the minimalistic model in Figure 4.5, one can see that the model outputs are in perfect accord with the time courses from BooleSim which further supports the filtering method applied in section 4.3. Furthermore, the stepwise progression in the different conditions corresponds a lot to what has been supposed in the explanation of the fits for the optimized models. In effect, it is SP1 that is switched off as initial actor as soon as KLF4 is present in the reprogramming cocktail, i.e. in the *KLF4*, *3TF* and *4TF* conditions inducing further steps such as the HIF1A and KLF4 down-regulation and thus also the re-silencing of *POU5F1* which was transitionally switched on by the combination of exogenous OCT4 and endogenous KLF4 in the latter two conditions. The CCND1 dynamics are a little more complicated. As long as c-MYC is absent from the retrovirally transduced gene cocktail, *CCND1* stays "inactive" as explained in section 4.3. When c-MYC is present as the

Figure 4.6: Rules Editor With Minimalistic Model for *4TF* condition

The general layout of BooleSim with the 3 different tabs *network*, *rules* and *time series* at the upper left and the options at the upper right. The JavaScript syntax for the BooleSim rule editor is shown with a grey background on the left and in the center the implementation of the optimized minimalistic model for early reprogramming is shown for the *4TF* condition where all retroviral genes are present, i.e. set to *true*

sole transduced factor, *CCND1* is expressed after 1 time step due to the activation by exogenous c-MYC and SP1 or exogenous MYC and endogenous MYC, the latter staying expressed in this condition because of the lack of retroviral KLF4. However, as soon as KLF4 is present as well, SP1 is down-regulated thereby inhibiting the cooperative activation of CCND1 together with c-MYC and thus an initial up-regulation of CCND1 is down-regulated again in the *4TF* condition.

In all 3 optimized networks from Section 4.3, all species in question are down-regulated or stay at low expression levels in the *3TF* and *4TF* conditions. This result was found as well when simulating the networks in BooleSim (results not all explicitly shown). However, the majority of targets are supposed to be active in pluripotent cells. Finding a way to prevent their down-regulation is thus strongly suspected to enhance the reprogramming process. I mentioned earlier that artificially keeping SP1 constitutively active could prevent the down-regulation of all the species. This could be confirmed in a simulation of all 3 networks when setting SP1 constitutively to *true* and independent of KLF4 regulation. It is interesting to notice that this permutation of the model would also make retroviral KLF4 redundant. I thus propose to test the transduction of a retroviral version of *SP1* together with the known cocktail or in combination with *POU5F1* or other single

transcription factors in order to enhance the process and replace at least KLF4. Another option would be to find a small molecule that keeps SP1 activated or to somehow "protect" the SP1 promoter from KLF4 binding and subsequent down-regulation.

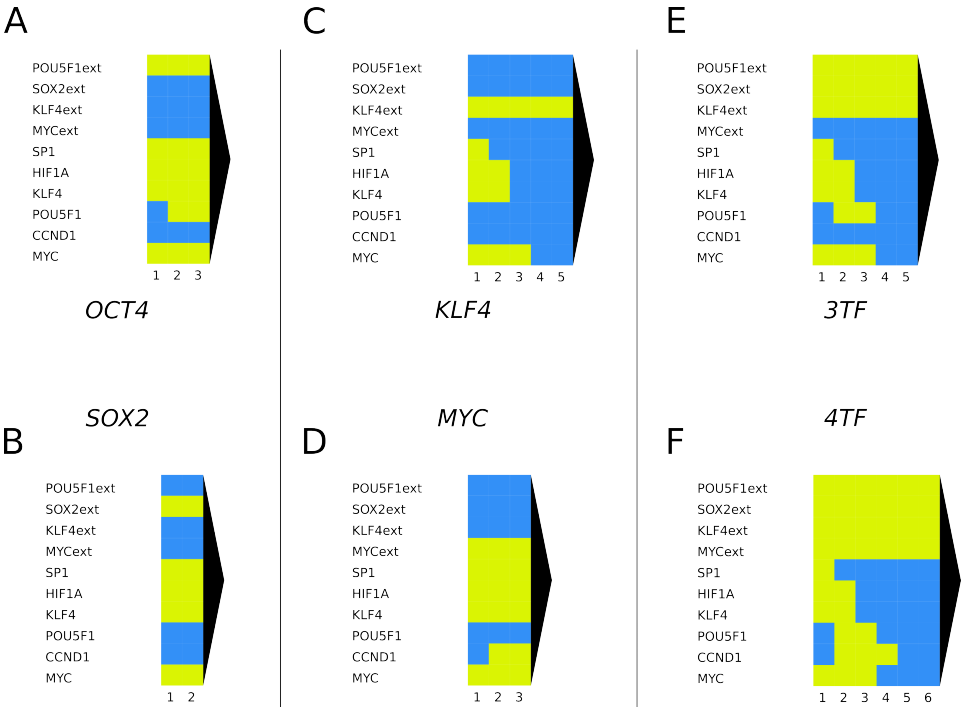

Figure 4.7: Time Courses of the Minimalistic Optimized Model

A In the *OCT4* condition every species remains as it is except for the endogenous *POU5F1* which gets activated. **B** In the *SOX2* condition, nothing happens which is due to the lack of connection of retroviral SOX2 to the rest of the consensus network and accounts for the lack of importance of SOX2 in the first 96 hours. **C** In the *KLF4* condition, the steady state after 96 hours consists of all species being shut off. The process happens gradually with SP1 being the first down-regulated species, followed by HIF1A and endogenous KLF4 and finally c-MYC while CCND1 and endogenous POU5F1 stay at a low expression level. **D** In the *MYC* condition, CCND1 is the only species changing its expression from 0 to 1. **E** The *3TF* combination of reprogramming factors shows the same behavior as the *KLF4* condition with the interesting difference that endogenous *POU5F1* is transiently activated but cannot sustain its expression. **F** The *4TF* condition shows the same behavior as the *3TF* condition with the difference that CCND1 is transiently activated.

By using it to analyze the time course of an optimized model, I have presented BooleSim, a Boolean modeling tool developed by us (Bock, Scharp, Talnikar, and Klipp, 2013) to provide users with a simple way to quickly gain qualitative understanding about the dynamics of the biological networks they want to study. It is the first Boolean modeling tool that includes an in-browser functionality while being able to work cross-platform between different formats. It refrains from the need to download a toole and enables users just via typing the websites URL http://rumo.biologie.hu-berlin.de/boolesim/ into a

browser window (Chromium recommended) to load, simulate and manipulate their Boolean networks. Moreover, BooleSim is open source and licensed under the free software license GNU Affero GPL version 3. The source code and an offline version of the tool are available for download on the GitHub repository (https://github.com/matthiasbock/BooleSim).

4.5 Summary and Discussion: Existence of a Transcriptionally Inactive Intermediate State?

In this Chapter, I have built a confident transcriptional interaction network from literature mining and intense expert curation making use of microarray gene expression profiling to filter for differentially expressed genes in the reprogramming process. An automated literature mining network with 41 differentially expressed genes and 295 interactions between them was thereby reduced to a highly confident network containing 26 genes and 75 mostly transcriptional interactions. This latter network is highly recommended to be used in the future for transcriptional modeling approaches. Network and data were then integrated using the CellNetOptimizer (CNO) framework described in detail in Terfve et al. (2012). Due to software constraints and methodological reasons, the network had to be further reduced to a final network containing 18 endogenous genes, the 4 retrovirally introduced genes and 53 interactions between these 22 species. The data and network were processed and the thus created set of models trained to the normalized data. The network processing was done as implemented in the software. However, the data processing, especially the normalization, had to be rethought following detailed analysis of the implemented procedures. As a consequence, data were manually rescaled between 0 and 1 and the model was trained against the thus created normalized data set. In the course of my research on the optimization of the network in question, I have repeatedly found that normalizing the data with respect to the *GFP* condition as control data at time point 0 yielded slightly better results than for the *FIB* condition, which supports the use of the former. Following this optimization procedure and a thorough analysis of the interactions that were found to be substantial in early reprogramming, the network was sensibly reduced by eliminating species that showed poor fitting to yield better optimization scores without losing a lot of downstream interactions and thus keeping the connectivity of the network strong.

While it was surprisingly found that the interactions between the master regulators of pluripotency, postulated by Boyer et al. (2005), don't seem to play a prominent role in early reprogramming, a new possible pathway of activation of endogenous *POU5F1* and *KLF4* was discovered that involves a

complex interplay of retroviral OCT4 and KLF4 together with endogenous SP1, IRS1 and STAT3. Since one of the factors at the base of this pathway, SP1, is down-regulated by retroviral KLF4, which was found to mark a very crucial step in early reprogramming, this mentioned pathway will possibly only play an activating role in later stages of reprogramming. Beside this effect, the down-regulation of SP1 has strong downstream consequences for FGF2 signaling, hypoxia response, cell cycle related CCND1 and even endogenous *c-MYC* expression which supports the hypothesis that SP1 is a crucial factor in early reprogramming.

The down-regulation of FGF2 could be reconciliated with the down-regulation of IRS1 through the PI3K pathway in order to keep MAPK/ERK at reasonable levels for pluripotent cells. However, the down-regulation of HIF1A and EPAS1 still leaves a few mysteries since their O_2-dependent regulation in general occurs at the protein level. The transcriptional down-regulation was thus explained as a consequence of the SP1 down-regulation and a possible intermediate state of hypoxia inducible factors. It is this SP1 down-regulation that also represses a possible CCND1 up-regulation observed in the *MYC* and *3TF* conditions via a complex mechanism involving exogenous or endogenous c-MYC, retroviral KLF4 and IRS1. A possible G_0/G_1 arrest in early reprogramming due to this repression could be hypothesized although it should be treated with care due to the still high expression values of *CCND1*.

Another result of the optimization algorithm consisted in the discarding of a few poorly fitted species, namely ID2, ID3, SMAD3, TGFBR2, PARP1 and GREM1. This could be due to a possible lack of involvement of these species in early reprogramming, to a lack of up-to-date knowledge about their transcriptional intertwining with the network or to the fact that their main mechanisms of action occur at a non-transcriptional level. Although constituting a negative result, it still contains the power to conclude that the involvement of the thus discarded species in the found transcriptional ways is questionable.

In fact, when taking together the results for the majority of the target genes included in the optimized models of pluripotency, it becomes clear, that although most of them are suspected to be strongly expressed and play an important role in iPSCs, they are all transiently down-regulated at 96 hours of early reprogramming. While a few explanations for this phenomenon have been tried in Section 4.3 including the crucial activation of hTERT, this early stage of reprogramming will definitely need more thorough attention and experimental efforts in the near future in order to understand the many controversies. However, as has been done with the intermediate state for the hypoxia inducible factors and the possible cell cycle arrest for CCND1, it is interesting to hypothesize the existence of an intermediate state in repro-

gramming that for one reason or the other shows low transcriptional activity of genes that need to be transcribed in later stages of reprogramming and in iPSCs. In fact, it is possible, that before these genes unfold their full transcriptional potential thus definitely determining the iPS cell lineage, they need to be held in suspense until different other re-structuring mechanisms have taken place. As we will see in Chapter 5, epigenetic re-structuring is one of these mechanisms that is directly related to transcriptional reprogramming but takes longer times itself to be fully in place. The hypothesis of the necessity of such an intermediate state should be tested by trying to surpass it in order to enhance reprogramming. If this turns out to be impossible, this would account for its necessity. One could imagine this intermediate state as a bow tensed with an arrow or a loaded spring that needs to be held in position until the aiming for the right target is finished before it can be released thus unfolding its full energy.

In addition to the experimental results, a few key features of the method that I used in this work should be discussed and their estimated influence on the results that were gained should be accounted for.

One of the first issues that always pops to the mind of the theoretical biophysicist is the problem of gene regulatory networks and microarray expression data in general. In fact, in this Chapter, I used data that describe the expression of a gene that can be correlated with the amount of transcript or even mRNA produced by the gene in question. However, the downstream action of the gene will be effected by the corresponding transcription factor that is the translated protein of the gene transcript and not by the mRNA. Therefore, one of the main assumption that is used in this optimization is the direct correlation between mRNA quantity and the corresponding protein concentration. Today it is known, that regulation takes place at the translation and protein processing levels as well defining protein concentration and activity via multiple mechanisms and that correlation between the two varies strongly from organism to organism. The experiments carried out in mammals thus far suggest a moderate correlation of approximately half of the tested proteins (Ghazalpour et al., 2011). However, modeling transcriptional regulatory networks with high-throughput data has been shown useful and is still widely practiced. Moreover, with the existing data and knowledge about networks involved in pluripotency, the method described in this Chapter and their results are a big step in the right direction of explaining mechanisms of early reprogramming and predicting novel targets for further experimental studies.

Another issue that needs to be addressed is the optimization of 2 time points only, one before and the other 96 hours after transduction with different combinations of reprogramming factors. Again, it should be noted, that experimental data in the field of reprogramming are still scarce although the

situation has improved recently. New data are expensive to reach and we have to work with what is existent. Having 2 data points that are measured with a certain temporal delay means that the tool will pretend that after 96 hours the system has reached a new steady state, so the steady state of the model will be compared to the data at 96 hours. Naturally, it is only visible what has happened after that time and not during that time. For example, if a species has been down-regulated after 96 hours, it could be, that it has first been up-regulated and then down-regulated or that it shows oscillatory behavior and we just look at it at a snapshot at 96 hours where it happens to be transiently down-regulated. Therefore, it is not possible to discriminate between more complicated dynamics when only considering the optimization of the model.

However, when implementing the optimized model in a simulation tool such as our BooleSim (Bock, Scharp, Talnikar, and Klipp, 2013), as was done in Section 4.4, it is possible to follow the dynamics of the found model. In this way, the order of events until the 96 hour time point could be retraced and it was especially found that the early KLF4-induced down-regulation of *SP1* is at the basis for the propagation of a signal that shuts down many of the down-stream target genes that are suspected to play an important role in the induction of pluripotency. It would be interesting to test whether this down-regulation is correlated with the low reprogramming efficiencies and whether preventing the *SP1* down-regulation by KLF4 can improve the process or whether on the contrary this is a necessary step - for example as hypothesized above to activate hTERT - that cannot be ignored. I therefore propose to carry out the reprogramming experiment with constitutively active *SP1* which is suspected enhance the reprogramming process and possibly to replace retroviral *KLF4* since it can activate endogenous *KLF4* on its own.

5 Stochasticity in Reprogramming: A Probabilstic Boolean Model Describing Transcriptional and Epigenetic Dynamics

The following Section is partly based on our publication: Flöttmann, Scharp, and Klipp (2012)

5.1 Epigenetics are Essential to Understand the Remaining Barriers

As outlined in the introduction, the regulation of cell differentiation takes place at a variety of different levels, e.g. transcriptional regulation, signaling pathways and various epigenetic processes such as DNA methylation and histone modifications to only name a few.

Up until now, I have only focused on transcriptional interaction networks in this work. Inside one cell line, where a certain set of housekeeping genes are prone to activation or inhibition, i.e. they are not inactivated by restrictive DNA methylation or heterochromatin structures, the analysis of such a transcriptional interaction network might be sufficient to explain a wealth of processes. However, when it comes to mechanisms that include epige-

netic restructuring such as cell (trans-)differentiation or reprogramming, it is necessary to include these layers of regulation into a model reflecting these processes. Moreover, although the analyses of network motifs and dynamics in transcriptional interaction models have helped to gain an insight into the importance of certain factors and structures, it still remains partly unclear why the reprogramming efficiencies, which are one of the main limiting factors of somatic cell reprogramming, are so low.

Therefore, in an innovative work, we created an abstract multilevel regulatory network, including transcriptional regulation between master regulators of pluripotency and 2 different cell lineages and epigenetic modifications of these genes, namely DNA methylation and histone modifications (Flöttmann, Scharp, and Klipp, 2012). I will now outline the main assumptions, methods and findings of this work and explain how they brought forward our understanding of the dynamics of reprogramming, differentiation and the interplay of epigenetics and transcriptional regulation.

As mentioned in the introduction, the successful reprogramming of somatic cells into induced pluripotent stem cells (iPS) (Takahashi and Yamanaka, 2006; Takahashi et al., 2007), has led transformation of cell types to become an important research field in recent years. Beside the reprogramming approach, it has thus been shown that the developmental state of a cell can be altered as well to transition between distinct differentiated cell types, the so-called trans-differentiation (Vierbuchen et al., 2010). In a more clinical perspective, there has been progress in autologous transplantation therapies in mice (Hanna et al., 2007), which however are still far from being used safely in human patients. In order to overcome the experimental hurdles and roadblocks (inefficiency, viral integration of oncogenes into the genome as mentioned in the introduction) on the way to patient specific clinical application of the reprogramming methods, it is necessary to improve our understanding of the exact mechanisms that underlie it.

In several attempts to enhance the efficiency and make the reprogramming approach more applicable for medicine, alternative techniques to the viral transduction of the 4 transcription factor cocktail proposed by Takahashi and Yamanaka (2006), have been developed. Apart from OCT4, all of the transcription factors in the cocktail are proto-oncogenes that will integrate into the genome upon viral transduction (Hochedlinger et al., 2005; Yancopoulos et al., 1985; Wei et al., 2006). Therefore, it was necessary to propose methods, that keep the genome unmodified or *signature-free*. These techniques include transfection with plasmids (Okita et al., 2008) that do not integrate into the genome or direct infusion of the transcription factor proteins encoded by the genes (Zhou et al., 2009). Although in theory improving the concept of clinical applicability, these methods are even less effective than classic reprogramming.

Other methods improving the efficiency include the addition of small chemical compounds (Wang and Adjaye, 2010), some of which can even replace the transcription factors KLF4 and c-MYC or even SOX2 depending on the cell lineage used and endogenous expression of the latter (Ichida et al., 2009) in the process. The majority of these small molecules have an influence on the epigenetic states and modifications of the cells that are responsible for the determination of the cell's developmental state. In this field of small molecules, valproic acid, a histone deacetylase 1 (HDAC1) inhibitor, has emerged as one of the most promising compounds improving reprogramming (Huangfu et al., 2008). It can be hypothesized, that inhibition of HDAC1 is capable of lowering the epigenetic barrier between different cell lineages making it easier for the cells to transition between different developmental states.

It was moreover shown that the reprogramming potential of a cell population is not restrained to specific cells in the culture as was hypothesized by critics of the Yamanaka publication. It is really rather the case that every cell can be reprogrammed and that due to the heterogeneity that exists even in one cell lineage, cells just need a different amount of time or more precisely a different amount of cell divisions per time unit, the so-called proliferation rate (Hanna et al., 2009). This finding is supported by the discovery that a high proliferation rate seems to act in favor of the reprogramming efficiency (Marión et al., 2009; Kawamura et al., 2009; Hong et al., 2009).

As partly outlined in the introduction, a cell's developmental state and thus also pluripotency is controlled by an interplay of regulatory mechanisms that take place at different molecular levels. We will outline 3 of these mechanisms which in reality are manifold, using the example of pluripotent stem cells. On the transcriptional level, proteins called transcription factors control the expression of target genes in a negative (inhibition) or positve (activation) manner. Master regulators of pluripotency or differentiation have the ability to regulate a wealth of downstream target genes thus determining the overall expression inside the cell lineage. In addition, we considered two layers of epigenetic regulation: DNA (de-)methylation of gene promoters and the structure of chromatin as active euchromatin or repressive heterochromatin. These 3 mechanisms will be outlined in more detail in the following.

The core transcriptional regulatory circuitry that accounts for pluripotency in human embryonic stem cells (hESCs), as discovered by Boyer et al. (2005), has been described in parts in the introduction in Subsection 1.2.1: OCT4, SOX2 and NANOG, the master regulators of pluripotency mutually induce their transcription and every transcription factor sustains its own expression. In a cooperative manner, they then activate a wealth of downstream target genes thereby promoting pluripotency and proliferation (Boyer et al., 2005; Loh et al., 2006). After its discovery, this 3-factor network was further extended in several studies generating different larger networks involved in

pluripotency (Ivanova et al., 2006; Zhou et al., 2007; Chavez et al., 2009).

DNA methylation has recently emerged as an important epigenetic regulatory mechanism which can silence gene promoters and thus control which genes are prone to transcription inside a cell lineage and which ones are not. Since reprogramming changes the developmental potential of a cell, the pluripotent cells having strongly different epigenetic marks than the differentiated cells, DNA methylation represents a major hindrance of direct reprogramming. This is because although active demethylation has been found to happen in reprogramming cells (Bhutani et al., 2011), in general DNA methylation cannot easily be reversed.

In recent years, due to the application of next generation sequencing techniques and the accumulation of molecular biological data, the so-called *methylomes* have emerged, i.e. the position of DNA methylation marks throughout the genome for different cell types (Lister et al., 2009; Laurent et al., 2010; Lister et al., 2011). It has become clear that while ESCs and iPSCs have similar DNA methylation states on a global scale, they strongly differ from somatic cells (Lister et al., 2011). However, it has also been found in the same publication that reprogramming of DNA methylation is very slow and that aberrant methylation sites remain in iPSCs, which partially accounts for the difference to ESCs. However, it was recently found that differential methylation of iPSCs and ESCs strongly diminishes over time (with continuous passaging of iPSCs) leading to a close resemblance of the two cell lineages. This process was shown to be driven by stochastic methylation and convergence of aberrant *de novo hyper-methylation* (Nishino et al., 2011).

Beside DNA methylation sites, comparative high-throughput studies also involved histone modifications that strongly affect the above mentioned chromatin structure. Taken together, a strong relationship between DNA methylation states and chromatin structure has been suggested (Hawkins et al., 2010). In fact, there seems to exist a substantial correlation between histone modifications involved in gene silencing and DNA methylations occurring inside the promoters of pluripotency master regulators (Cedar and Bergman, 2009). How this correlation works exactly on the molecular level still remains partially unclear. At the basis of this mechanism are probably histone binding proteins with a histone methylation activity conveyed by the histone methyltransferase (HMT). The specific methylation of histones can facilitate the formation of repressive or activating heterochromatin depending on the histone protein and amino acid residue in question.

Alongside its HMT activity, G9a also recruits the DNA *de novo* methyl transferases DNMT3A and DNMT3B to the nucleosome which are able to methylate the DNA, especially the promoter of genes. In addition to the gene expression regulating features of DNA methylation, it is also suspected to help stabilize chromatin structures during mitosis via the binding of different

proteins for closed (hetero-) or open (eu-)chromatin (Cedar and Bergman, 2009). Moreovoer, it is thought to inhibit triple methylation of residue lysine 4 at histone 3 (H3K4me3), an activating histone mark. The *epigenetic memory* that describes the inheritance of certain histone modifications onto the daughter strand after mitosis is coupled to the DNA methylation pattern as it guides binding of HDACs (Fuks et al., 2000).

As for DNA methylation, it is passed onto the next generation throughout DNA replication and mitosis thanks to DNMT1 by directly reproducing the methylation pattern of the template mother strand onto the copied daughter strand. Although this process is very efficient, methylation marks can still be lost in rapidly proliferating cells and cells lacking DNMT1 (Monk et al., 1991).

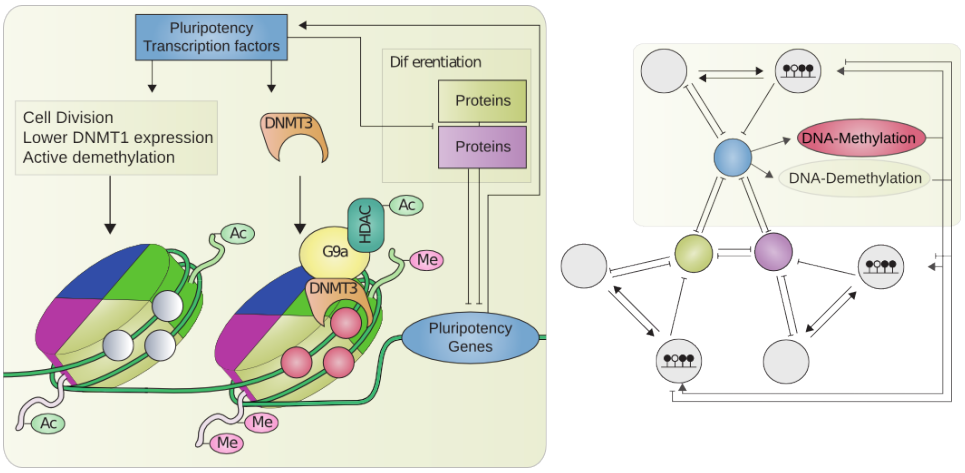

Figure 5.1: Molecular Mechanisms Underlying the Model Derivation and Schematic Model Representation (taken from Flöttmann, Scharp, and Klipp (2012))

A: The molecular interplay between DNA methylation, histone modifications and transcription of the module including pluripotency master regulators is shown. The pluripotency TFs sustain their own expression and are inhibited by master regulators of other differentiated cell lineages. The pluripotency TFs then positively act on *DNMT3* expression, cell proliferation, demethylation and suppress *DNMT1* expression leading to decreased methylation maintenance. DNMT3 counteracts this latter effect by recruiting G9A and HDAC to the nucleosome leading to decreased histone acetylation and increased DNA and histone methylation leading to repressive chromatin. **B** The schematic structure of our PBN model without the external module. A more detailed description of it is shown Figure 5.2

As described in the introduction in Section 1.3, several mathematical models have been established to approach the regulation of pluripotency and somatic cell reprogramming. These approaches were mainly focused on one specific regulatory feature such as transcriptional interactions (Chickarmane and Peterson, 2008; MacArthur et al., 2008). However, if we focus on one part of a big system only, the complex interplay of the different parts that

constitute a cell and that are crucial to its functioning, can never be analyzed. Therefore, we built an abstract, holistic model in order to be able to combine transcriptional regulation and two different epigenetic mechanisms. In Figure 5.1, a schematic representation of the interplay of the molecular mechanisms (A) and their abstraction in a regulatory graph (B) is shown.

Our abstract model includes 3 layers of interfering regulatory mechanisms that control pluripotency and reprogramming. We are adopting a recently developed modeling framework, probabilistic Boolean networks (PBNs), in a new conceptual manner (see Figure 5.2) in order to reflect the clearly non-deterministic nature of the processes involved. Since PBNs are based on a standard Boolean networks approach and since our model is built in a modular structure, it can easily be changed, extended and merged with results from other Boolean approaches. Since in Boolean models states can be represented as a bitstring or a binary vector of 0s and 1s it is very easy to compare states with each other. We will derive the exact model strucure in Section 5.3.

5.2 Probabilistic Boolean Modelling as a Way to Handle Uncertainty in Epigenetic Modeling

There are different ways to introduce stochasticity into Boolean networks, e.g. via asynchronous updating, stochasticity in nodes (SIN) or stochasticity in functions (SIF) and probabilistic Boolean networks (PBNs) (Harvey and Bossomaier, 1997; Garg et al., 2009; Shmulevich, 2002; Twardziok et al., 2010). In this work, we will focus on the probabilistic Boolean network approach proposed by Shmulevich (2002) and described in detail in Section 2.3.4.

In our approach, we use PBN modelling to account for two different kinds of stochasticity. The first is the uncertainty on the level of the Boolean functions arising from the lack of knowledge of exact molecular mechanisms and data. Using the probabilistic approach, we can try different variants to find out which possible underlying mechanisms might reproduce literature findings best.

The second way in which the probabilistic approach fits our needs is to include the stochastic features of transcription and epigenetics that we want to model. By assigning different functions with varying probabilities, we can try to reconstruct the stochastic nature of biological processes.

Apart from the inherent stochastic nature of PBNs, their simulation can also be effected in a stochastic or a "deterministic" manner. In the former way single trajectories of the model are simulated a certain number of times and

the results are analyzed like the outcome of a stochastic experiment. This is to say, one chooses a certain CBN model out of the ensemble of CBN models constituted by the PBN with the underlying probability and runs the simulation of this CBN. This process is carried out a substantial amount of times and the results are averaged yielding a probability distribution of each state over time. In a more deterministic and mathematically more challenging manner, it is also possible to analyze the resulting discrete Markov chain, which will be our mode of choice which is explained in detail in Subsection 2.3.4.

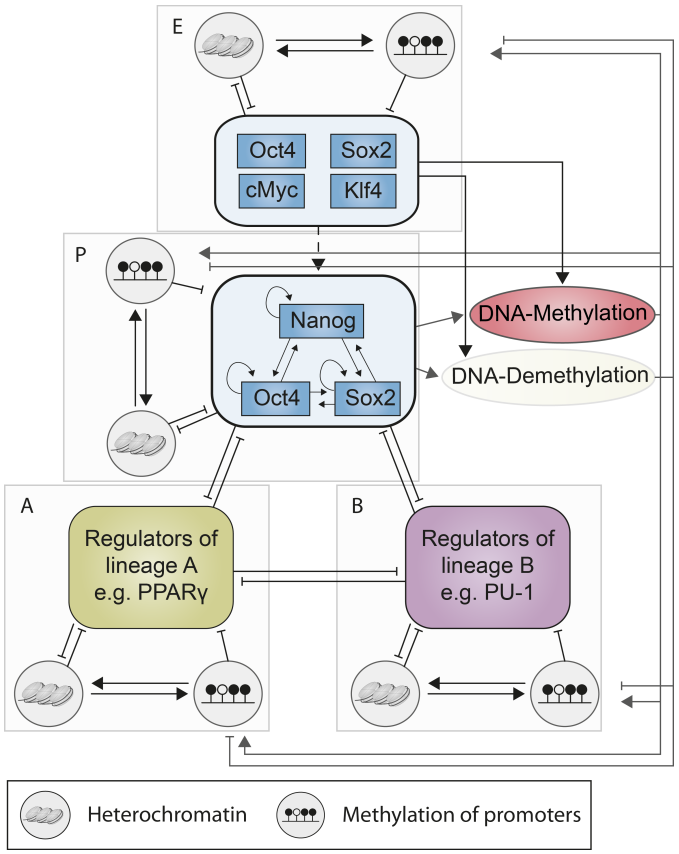

Figure 5.2: General Model Structure of the Complete PBN (taken from Flöttmann, Scharp, and Klipp (2012))

Our complete model is built of 4 modules representing the retrovirally introduced genes (E), the pluripotency master regulators (P), a differentiated cell lineage (A) and another cell lineage (B). Each module is composed of 3 species accounting for a transcriptional activity of the master regulators of the module, a chromatin state and a DNA methylation state. The species inside of a module influence each other in the shown way. Moreover the transcriptional activity of a module influences the one of other modules as well. Furthermore, a DNA methylation and a DNA demethylation activity is also part of the model.

5.3 Derivation of the Model

When trying to build a model with high regulatory complexity containing processes that take place at different places inside a cell and with various timings for the first time, a high level of abstraction is inevitable. Therefore, we chose to modularize the structure by summarizing several similar factors into one module behaving like one distinct species and similar processes into interactions between these modules. This approach was used by modellers before and has shown to generate coherent results (Artyomov et al., 2010).

We constructed our model of 4 big modules, one representing the exogenous factors (module (E) in Figure 5.2), one representing the endogenous pluripotency factors (module (P) in Figure 5.2) and two modules representing two distinct cell lineages (modules (A) and (B) in Figure 5.2). Every one of these 4 modules consists of a transcriptional part, a DNA methylation state and a chromatin species. It can be interpreted as a group of genes governing the morphology and function of the cell, i.e. the specificity of the cell lineage, the general DNA methylation pattern of these genes and the situation of these genes inside transcriptionally active or inactive chromatin due to histone modifications. These 4 modules have very similar but not exactly the same structure as will be explained in more detail below. In addition to these 4 big modules, we added 2 species governing DNA methylation (Called *dnmt* in our model) and DNA demetyhlation (called *demeth* in the model).

The transcriptional species inside a module contains activating transcriptional interactions between its members. As such, the network responsible for sustaining pluripotency consisting of OCT4, SOX2 and NANOG has been explained in detail in the introduction in Subsection 1.2.1. They form a transcriptional circuitry that is mutually and auto-activating (Boyer et al., 2005). For differentiated cell lineages, similar structures of interacting transcriptional master regulators have been found, such as PU-1 in erythrocytes (Nishimura et al., 2000; Okuno et al., 2005) or PPARγ in adipose tissue (Wu et al., 1999).

Between these modules, i.e. between different transcriptomes of cell lineages, interactions are often mutually repressive, e.g. GATA-1 and PU-1 (Rekhtman et al., 1999). The pluripotency module also represses differentiation factors as has been modelled for instance in Chickarmane and Peterson (2008). This mutual antagonism paired with auto-activation of the single modules is the basic structure of the transcription factor regulations in our model.

On top of the transcriptional interactions inside and between modules, we have also included 2 different epigenetic features that influence each other and the gene expression of the transcriptional part and that will be described

in the following. The main concepts that we have followed when deriving the interplay between DNA methylation and chromatin formation due to histone modifications is based on Cedar and Bergman (2009).

Epigenetic marks define a higher and more permanent level of regulation than transcriptional interactions. In fact, the epigenetic state of the cell defines which genes can be transcribed when the transcription factor machinery is recruited to their promoter and which ones are in a restrictive environment which disallows them to be accessed by transcription factors and RNA polymerases. Moreover, epigenetic marks are made to be more or less permanent in order for terminally differentiated cells not to trans-differentiate spontaneously into other cell lineages.

The expressed transcription factors, signaling proteins and RNAs, beside determining future expression profiles through action on target gene promoters, also affect the epigenetic marks. The latter then in turn define a new cellular transcriptome and thus proteome. This mutual interplay is one of the basic assumptions generating the internal structure of our model. In our model, expression of the transcription factors of one module, favors the removal of restrictive chromatin marks and there is a certain probability to also remove repressive DNA methylation marks. On the other hand, the silencing mechanism, i.e. DNA methylation and heterochromatin formation, is possible to happen, when the genes of the module are not expressed. Since DNA methylation and histone marks favoring heterochromatin mutually enhance each other (Epsztejn-Litman et al., 2008; Thomson et al., 2010), this is reflected in the model as well

In biological reality, DNA methylation can occur at many different CG dinucleotides upstream, inside or downstream of a gene to act on its expression. As for the transcriptional part of the module, we suppose the many DNA methylation sites of one gene to be highly co-regulated in order to be able to model the overall DNA methylation state of the gene as one species in our Boolean model. Hence, the entity associated with it can either be active (methylated) leading to lack of expression of the genes of the module or inactive (demethylated) leading to possible transcription. As mentioned above, at the base of the interplay with histone modifications and responsible for the transfer of new methyl groups onto the DNA are the *de novo* methyltransferases DNMT3A/B. These entities are summarized in the variable *dnmt*.

Methylated DNA can also be demethylated by various mechanisms. During cell replication, the newly created strand of DNA is not methylated at first and will only be methylated via an active DNMT1 whose inefficiency or failure can account for passive demethylation (Monk et al., 1991). Furthermore, recent discoveries show that there might be active demethylation patterns as well (see table 5.1 and Ou et al. (2007)). These demethylation processes are

summarized in the variable *demeth* in our model. Epigenetic processes such as DNA methylation and demethylation occur at much slower paces than transcriptional changes. To account for this in our model, we introduced an update rule leaving the DNA methylation state as it is with a high probability. All these findings taken together result in the following Boolean update rules for DNA methylation of modules (A), (B) and (P):

$$\begin{aligned} m_m^A(t+1) &= m_m^A(t) \vee dnmt(t) \wedge m_{hc}^A \\ m_m^A(t+1) &= m_m^A(t) \wedge (demeth(t) \vee m_{hc}^A) \\ m_m^A(t+1) &= m_m^A(t) \wedge demeth(t) \\ m_m^A(t+1) &= m_m^A(t) \end{aligned} \tag{5.1}$$

where m_m^A and m_{hc}^A are the methylation and chromatin states of module A, respectively. The *dnmt* and *demeth* variables are governed by the following rules:

$$\begin{aligned} dnmt(t+1) &= m_e^P(t) \vee m_e^E(t) \\ dnmt(t+1) &= m_e^P(t) \vee m_e^E(t) \vee dnmt(t) \\ demeth(t+1) &= m_e^P(t) \vee m_e^E(t) \\ demeth(t+1) &= m_e^P(t) \vee m_e^E(t) \vee demeth(t) \end{aligned} \tag{5.2}$$

where m_e^P and m_e^E represent the expression of the pluripotency and the exogenous modules, respectively. The probabilities associated with the update function containing the species itself (the 2nd and the 4th in Equations 5.2) are very high (see table 5.1) while the other 2 are very low, i.e. turning off these factors is slow. We introduced this feature because on the one hand we assume that these are not the only influences on these variables and that they need to be active in many cell states and on the other hand we include a stochastic equilibrium between methylation and demethylation which might lead to interesting dynamics.

As for the other parts of the modules, the histone modifications as well are greatly simplified in our model. We don't consider neither the type of modification nor the quantity of modifications made. Just as for the DNA methylation, we only consider transcriptionally active or inactive chromatin and factors that favor the one or the other. Chromatin changes are dependent on the expression of the module's genes. When the genes of a module are not expressed, there is a chance of repressive histone modifications to form which is further favored by DNA methylation marks (Feldman et al., 2006;

Cedar and Bergman, 2009). In Boolean formulas the above discussed looks as follows:

$$\begin{aligned} m_{hc}^A(t+1) &= m_{hc}^A(t) \vee m_m^A(t) \wedge \neg m_e^A(t) \\ m_{hc}^A(t+1) &= m_{hc}^A(t) \vee \neg m_e^A(t) \\ m_{hc}^A(t+1) &= m_{hc}^A(t) \wedge \neg m_m^A(t) \\ m_{hc}^A(t+1) &= m_{hc}^A(t) \end{aligned} \tag{5.3}$$

where m_e^A is the variable representing the expression of module A, m_{hc}^A the chromatin state and m_m^A the DNA methylation of the module respectively. According to these rules, present DNA methylation marks increases the probability of heterochromatinization. As we have seen above, the same holds for the dependence of methylation on the chromatin state of the module. Thus, these Boolean formulas reflect the mutually enhancing structure of DNA methylation and heterochromatin formation that has been mentioned several times before.

Concerning the expression of a module's genes, it is governed by its epigenetic states since chromatin and DNA methylation have a strong influence on gene expression. If the gene is located in heterochromatin and methylated it is completely silenced and cannot be activated by transcription factors anymore. In the case where both epigenetic marks are not set, the genes are prone to expression if transcriptional activators are present as it would be in a purely gene regulatory network. If only one of the marks is set, transcription of the corresponding genes is possible with a lower probability. This behavior is reflected in the following Boolean rules and is the same for all modules:

$$\begin{aligned} m_e^A(t+1) &= m_e^A(t) \wedge \neg(m_e^B \vee m_e^P(t)) \wedge \neg m_m^A(t) \\ m_e^A(t+1) &= m_e^A(t) \wedge \neg(m_e^B \vee m_e^P(t)) \wedge \neg m_{hc}^A(t) \end{aligned} \tag{5.4}$$

For the first time since the discovery of reprogrammin, we also modeled the exogenous viral factors and their action on the endogenous pluripotency genes. With a low probability, these exogenous factors can activate the endogenous pluripotency network. When the reprogramming process is over, i.e. the endogenous pluripotency module is active while the modules for differentiation are turned off, the viral vectors are silenced by epigenetic marks (as reviewed in Hotta and Ellis (2008)). The reason for the low probability of activation of the pluripotency module by the exogenous factors lies in the fact that only 4 factors are transduced while the whole ensemble of

pluripotency governing factors is made up by a wealth of genes. Therefore the probability of the endogenous pluripotency module sustaining its own activity should be much higher.

As explained earlier, the viral gene duplicates have a different promoter region than their endogenous pluripotency equivalents. Therefore, the exogenous module will behave differently on the transcriptional level since it is not regulated by any endogenous factors but only by their epigenetic state. However, the regulation of the latter will also be modified in comparison to the other modules. For the the viral factors' gene expression, the above yields the following equations:

$$\begin{aligned} m_e^E(t+1) &= m_{hc}^E(t) \vee m_m^E(t) \\ m_e^E(t+1) &= m_{hc}^E(t) \wedge m_m^E(t) \end{aligned} \tag{5.5}$$

The rules for methylation of the promoter of the exogenous genes are very similar to the ones of the other modules except for the probabilities which we chose to be smaller for *dnmt* and heterochromatin dependent DNA methylation. In fact, after reprogramming, it is possible to observe cells where the retroviral genes are still expressed (the so called class I iPSCs) while in others they are epigenetically silenced and thus fully reprogrammed (called class II iPSCs) (Niwa, 2007b; Mikkelsen et al., 2008). These incomplete methylation patterns, combined with the fact that DNA methylation doesn't seem to be needed to abolish retroviral gene expression (Pannell et al., 2000) justify these low probabilities.

In the same way as for the other modules, for the exogenous module as well there is slow (low probability of change), cell cycle dependent DNA demethylation, which might be due to variable activity of DNMT1 after mitosis (Li et al., 1992) (also see Table 5.1).

All other update rules for DNA methylation are the same as for the other modules. Thus, the structural difference is summarized in the following rule:

$$m_m^E(t+1) = m_m^E(t) \wedge (\neg demeth(t) \vee dnmt(t)) \tag{5.6}$$

When it comes to chromatin modification rules of the retroviral genes, we included one of our hypotheses that distinguishes retroviral silencing from the epigenetic silencing of the other modules. In fact, there needs to be a mechanisms that takes into account the timing of reprogramming because retroviral silencing only takes place in fully reprogrammed iPSCs. Moreover, this mechanisms needs to be independent of DNA methylation (Pannell et al.,

2000), in contrast to the epigenetic crosstalk of other modules. We hypothesized that the NANOG and OCT4 associated deacetylase (NODE) complex or a complex with similar characteristics is responsible for this mechanism. It is constituted by a histone deacetylase (HDAC) and NANOG or OCT4 (Liang et al., 2008) and was found to catalyze histone deacetylation on developmental target genes thereby leading to heterochromatin formation (Hotta and Ellis, 2008). Due to the fact that the complex needs NANOG or OCT4, the corresponding update rule, which is the only one that structurally differs from the other modules on the chromatin level, depends on the expression of the pluripotency module P:

$$m_{hc}^{E}(t+1) = m_{hc}^{E}(t) \vee m_{e}^{P}(t) \tag{5.7}$$

We have now listed the complete set of update rules constituting our model. For a summary of the update rules and a visual representation of the general model structure as outlined above, please consult Table 5.1 and Figure 5.2.

Table 5.1: General PBN Model Structure With Literature Evidence:

In bold in the **Update Rule** column, we represent the part of the variable's update rule that reflects the modeled property described in column **Represented Property** and further explained and literature referenced in column **Explanation**. The column **Probability** contains the probabilities of the update rule

Represented property	**Update Rule**	**Probability**	**Explanation**
Auto activation of gene modules	$m_e^A(t+1) = \mathbf{m_e^A(t)} \wedge \neg(m_e^B(t) \vee m_e^P(t)) \wedge \neg m_{m/hc}^A(t)$	0.5/0.5	Regulatory proteins are closely co-regulated and are often connected by positive feedback loops. (Boyer et al., 2005; Chickarmane and Peterson, 2008; MacArthur et al., 2008)
Pluripotency module activating DNA methylation through variable *DNMT* expression	$dnmt(t+1)/demeth(t+1) = \mathbf{m_e^P(t)} \vee m_e^E(t) \vee dnmt(t)/demeth(t)$	0.99	*DNMT3* co-regulated with Pluripotency genes. DNMT3 methylates unspecifically (Adewumi et al., 2007; Mah et al., 2011). Processes that contribute to DNA demethylation are regulated in the same manner as DNMT3 in our model, further introducing a stochastic equilibrium between the two processes.
Mutual inhibition of gene modules	$m_e^A(t+1) = m_e^A(t) \wedge \boldsymbol{\neg(m_e^B(t) \vee m_e^P(t))} \wedge \neg m_{m/hc}^A(t)$	0.5/0.5	Master Regulators inhibit other master regulators, competing lineages repress each other (Niwa et al., 2005b; Ralston and Rossant, 2005; MacArthur et al., 2008)
Heterochromatin increases probability for DNA methylation	$m_m^A(t+1) = m_m^A(t) \vee dnmt(t) \wedge \mathbf{m_{hc}^A(t)}$	0.05	Interaction via G9a complex: DNMT3A/B bind to nucleosomes with methylated histones such as H3K9me and methylates DNA (Cedar and Bergman, 2009)
Heterochromatin formation is inhibited by appropriate gene module	$m_{hc}^A(t+1) = m_{hc}^A(t) \vee m_m^A(t) \wedge \boldsymbol{\neg m_e^A(t)}$	0.11	G9a binds specific sequences (Epsztejn-Litman et al., 2008)

DNA methylation increases probability for heterochromatin formation	$m_{hc}^{A}(t+1) = m_{hc}^{A}(t) \vee \mathbf{m_m^A(t)} \wedge \neg m_e^A(t)$	0.17	Promotes chromatin inheritance after mitosis (Thomson et al., 2010)
DNA demethylation slower than other factors	$m_m^A(t+1) = m_m^A(t) \wedge \mathbf{demeth(t)}$	0.02	Passive cell cycle dependent demethylation through variable DNMT1 activity after mitosis (Li et al., 1992)
DNA demethylation is faster in euchromatin	$m_m^A(t+1) = m_m^A(t) \wedge (\mathbf{demeth(t)} \vee \mathbf{m_{hc}^A})$	0.03	Histone deacetylase (HDAC) inhibitor TSA induces global and specific DNA demethylation (Ou et al., 2007)
Methylation not necessary to downregulate retroviral gene expression	$m_e^E(t+1) = \neg m_{hc}^E(t) \neg \vee \boldsymbol{m_m^E(t)}$	0.5	Retroviral silencing is DNMT3A/B independent in the first 10 days of reprogramming (Pannell et al., 2000)
Retroviral gene demethylation is very slow in absence of DNMT3A/B or DNMT1	$m_m^E(t+1) = m_m^E(t) \wedge \boldsymbol{(\neg demeth(t) \vee dnmt(t))}$	0.001	
Retroviral gene heterochromatin dynamics	$m_{hc}^E(t+1) = m_{hc}^E(t) \vee \boldsymbol{m_e^P(t)}$	0.1	A complex between HDAC and NANOG (NODE complex responsible for the silencing of developmental genes) could account for retroviral silencing (Hotta and Ellis, 2008; Liang et al., 2008)

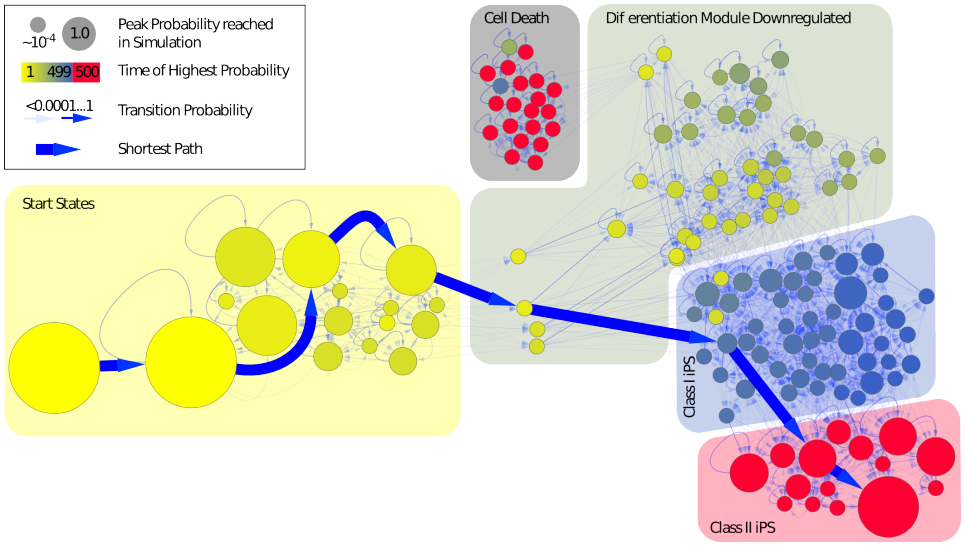

Figure 5.3: State Space and Reprogramming as a Path Through it (taken from Flöttmann, Scharp, and Klipp (2012))

The Figure shows the 149 out of 16384 states of the state space that are reached with a minimum probability of $p \geq 10^{-4}$ in a time course simulation over 500 time points of a reprogramming experiment. The legend in the upper left explains the color and size codes of the nodes and edges. Taken together, different phases of reprogramming can be discerned including an epigenetic modification phase (yellow), followed by a phase in which the transcriptional activity of the differentiation related module is down-regulated. From this phase, cells can either transition to undefined states that can be related to cell death or to the class I iPSCs (blue) in which the pluripotency master regulators are expressed. From the class I iPSCs cells then transition to class II iPSCs (red) in which the retroviral genes are silenced. While many paths with different length exist for the reprogramming process, the fastest path includes 7 transitions and is marked by the thick blue arrows.

5.3.1 Simulations of a Single Module

Before starting to simulate our complete model that consists of all 4 modules mentioned earlier, we start by analyzing some of the single modules on their own to deduce their behavior as standalone models.

As mentioned before, each of our modules is built up of 3 parts, namely the gene expression, the DNA methylation state and the chromatin structure. The DNA methylation state is regulated by the modifiers *dnmt* and *demeth* (as summarized in Figure 5.2). There is an inherent difference between modules A and B that are responsible for differentiation and maintenance of their cell lineages and the pluripotency module P responsible for pluripotency. While the former only regulate their own state and repress expression of other modules, the pluripotency module additionally influences the expression of *dnmt* and *demeth*. Therefore, the behavior of these 2 parts is essentially different.

Without any external influences in the standalone pluripotency module, the state in which the pluripotency genes are active is stable. Artificially converting the the chromatin state to heterochromatin yields partly silencing but also partly a return of the expressed state. Upon DNA methylation, the pluripotency genes are completely silenced and the chromatin state is locked. Constantly expressing a transcriptional repressor of the pluripotency genes (e.g. master regulators of cell lineage A or B) yields transcriptional silencing of the pluripotency genes and a dynamic equilibrium between states that include heterochromatin marks and active or inactive *dnmt* (Figure 5.5).

As for the pluripotency modules, the differentiation related modules A and B are stable as well if no other genes or external factors are expressed. Just as for the pluripotency example, if the genes of the other cell lineage are expressed, the differentiation module is transcriptionally silenced and its heterochromatin state fluctuates because there is no DNA methylation (Figure 5.4 B).

However, if the pluripotency genes are expressed in the differentiated state, the dynamical behavior is. The situation resembles a strongly simplified reprogramming experiment. Obviously, the gene expression of the differentiation is repressed by the constantly expressed pluripotency genes. Moreover, the epigenetic marks enter an equilibrium fluctuating between different states (Figure 5.4 A). This hyperdynamic plasticity has been observed in differentiation genes in pluripotent cells and described by Niwa (2007b). Through the action of the epigenetic modificators, changes in DNA methylation states are induced leading to a high probability of module A to have methylated DNA marks. Upon deactivation of the pluripotency signal, the system does not reverse its behavior completely and return to the start state, but is partially arrested in non-physiological undetermined states without expression of any module.

Although reprogramming experiments seem to be easy to implement and simulate in the Section above, we can still not answer the question concerning the low reprogramming efficiency and we still haven't considered neither the interplay of the complete 3 modules mentioned above neither the external factors needed for reprogramming. We thus combined modules A, B and P in a preliminary model before including the retroviral genes of module E, which have a regulation of their own and a completely different influence on the model.

5.3.2 Stable Cell States and Differentiation of Combined Modules

The above mentioned combination of modules A, B and P (3 out of the 4 modules of Figure 5.2 without the external factors) yields a more complex dynamic behavior. From the network structure, it is clear, that gene expression in one module is mutually exclusive with all other modules, i.e. on the long run, only one module's genes can be stably expressed and the system has to migrate into one or the other set of states. A module whose expression is deactivated can only be expressed again by external influence together with epigenetic re-modeling. The active pluripotent state, i.e. the steady state of module P consists of a distribution of several similar states that account for the hyperdynamic plasticity of epigenetic marks observed in pluripotent cells (Meshorer et al., 2006). In this distribution, depending on the exact epigenetic configuration, states have different probabilities to re-differentiate, a feature that has been found in populations of pluripotent cells regarding the expression of *NANOG*, one of the master regulators of pluripotency (Kalmar et al., 2009).

Since our model mainly focuses on epigenetics and transcriptional interactions and thus already includes 3 different mechanisms in one very simplified model, we neglect the action of signalling pathways which also have a substantial influence on differentiation processes. For the simulation of differentiation, we activate gene expression of the differentiation module in question which leads to quick deactivation of the pluripotency genes and after approximately about 300 time steps to a differentiation related steady state (Figure 5.7 A). Furthermore, as observed in the single modules as well, the system partly gets stuck in an undetermined state, in which all genes are unexpressed. A possible explanation for this intermediate, undesired state is that we strongly simplified the differentiation signal which could lead to an uncontrolled timing of events: For example, if the pluripotency module is deactivated before the epigenetic pattern of the differentiation modules was set up properly, this could lead to this undetermined state since genes of the differentiated cell lineage cannot be expressed due to the epigenetic structure. At the same time, pluripotency genes could already have been deactivated and thus *de novo* DNA methylation and DNA demethylation mechanisms could be inactive leading the system into a dead-end steady state. A proper setup of signalling pathways, which we refrain from as mentioned above, could prevent this behavior.

However, it is possible as well, that instead of being due to simplification, this behavior reflects de-regulation that also occurs in biological systems. In biochemical reality, it can be caused by transcriptional noise, epigenetic variability, or different external factors. In this way, the undefined state could

be related to cell death, such as necrosis or apoptosis caused by the applied stimulus. Altogether, although the model is strongly simplified, it is already able to quickly differentiate out of a pluripotent cell lineage into stable differentiated states thus reconstructing real differentiation experiments as shown in Table 5.3.

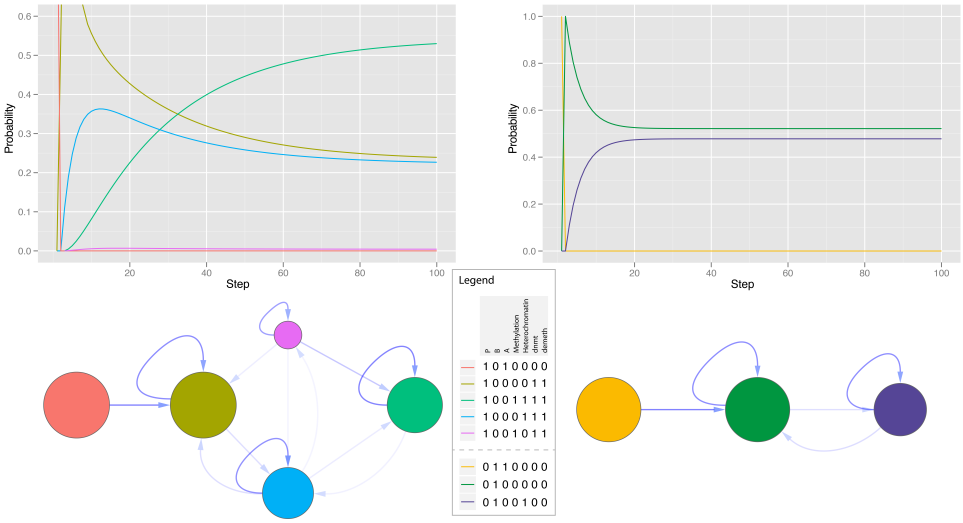

Figure 5.4: Time Courses of Single Modules of Differentiated Cell Lineages (taken from Flöttmann, Scharp, and Klipp (2012))

A Time course of the active differentiated cell lineage module (A) with a fixed value of 1 for the transcriptional activity of the pluripotency module (P) and of 0 for the cell lineage (B). We can observe down-regulation of the transcriptional activity of (A) and up-regulation of *dnmt* and *demeth* followed by the appearance of heterochromatin and methylation marks. **B** If the pluripotency module is inactive and cell lineage (B) is transcriptionally activated, we observe down-regulation of transcriptional activity of (A) and possible heterochromatin formation which is not stable however due to a lack of DNA methylation. It should be noted that the time courses show the change of the probability of the network to be in the state (corresponding to the color) over time.

5.4 Integrating Retroviral Reprogramming Factors

Finally, to analyze the reprogramming process within our full model framework, we combined the four single modules, i.e. the retroviral transcription factors E, the endogenous pluripotency genes P and the two model cell lineages A and B into one model of reprogramming and differentiation (Figure 5.2). We ran a Markov simulation of the whole model for various starting distributions and qualitatively analyzed the dynamics of the model for typical experimental scenarios.

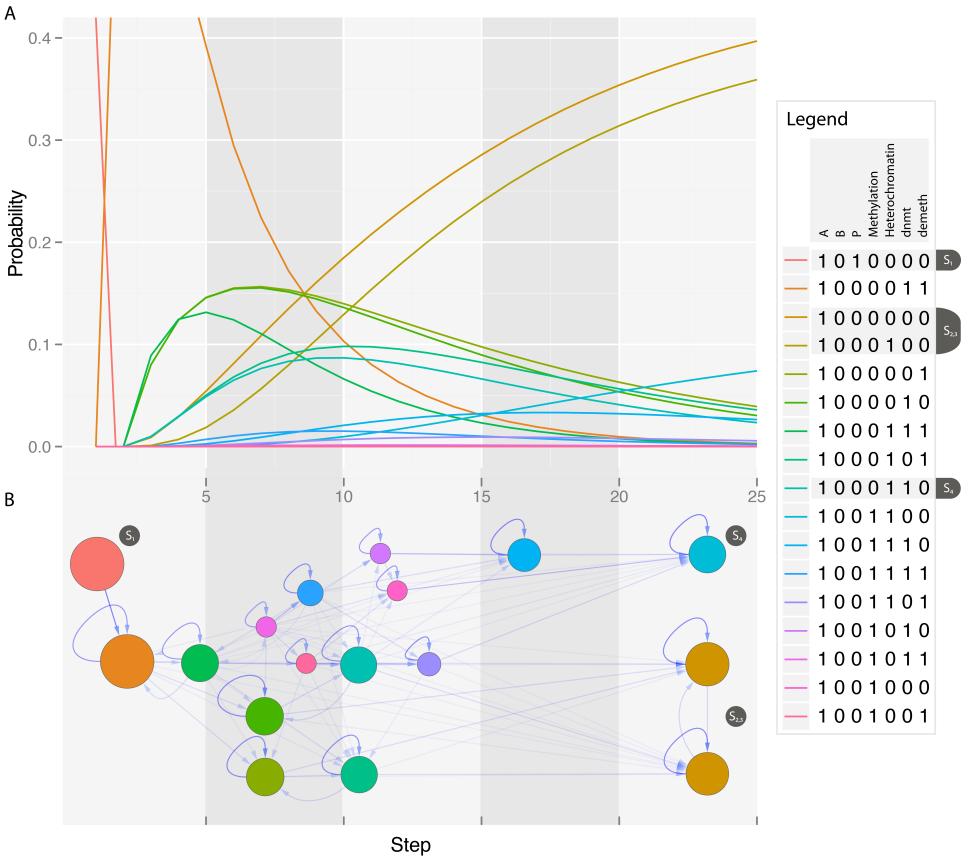

Figure 5.5: Differentiation Time Course of Single Pluripotency Module (taken from Flöttmann, Scharp, and Klipp (2012))

When constitutively expressing the master regulator of one of the differentiated cell lineages, the network quickly leaves the pluripotent state and transitions into differentiated states with different epigenetic marks of the pluripotency module (States $S_{2,3}$ and S_4)

Table 5.2: Variables and states of our model (taken from Flöttmann, Scharp, and Klipp (2012))

The columns represent the model's variables. In the rows, the pluripotent and the two differentiated states as Boolean states as well as the weight vectors explained in Section 2.3.5 and used for the state sorting in Figure 5.7 are shown

	m_e^E	m_m^E	m_{hc}^E	m_e^P	m_m^P	m_{hc}^P	m_e^A	m_m^A	m_{hc}^A	m_e^B	m_m^B	m_{hc}^B	$dnmt$	$demeth$
Pluripotent state $\mathbf{S_1}$	0	1	1	1	0	0	0	1	1	0	1	1	1	1
Differentiated state $\mathbf{S_2}$	0	1	1	0	1	1	1	0	0	0	1	1	0	0
Differentiated state $\mathbf{S_3}$	0	1	1	0	1	1	0	1	1	1	0	0	0	0
Weightvector $\mathbf{W_1}$	0.5	0.5	0.5	2.0	10.0	5.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0
Weightvector $\mathbf{W_1}$	0.5	0.5	0.5	2.0	2.0	2.0	2.0	10.0	5.0	2.0	2.0	2.0	1.0	1.0
Weightvector $\mathbf{W_1}$	0.5	0.5	0.5	2.0	2.0	2.0	2.0	2.0	2.0	2.0	10.0	5.0	1.0	1.0

To test the stability of the model's cell lineages, we set the system's initial conditions, i.e. the starting state of the simulation in our first analysis to correspond to either one of the two cell lineages A and B. In this state, the set of master regulator genes associated with lineage A is expressed, its DNA unmethylated, and the genes are in an open, transcriptionally prone chromatin configuration. The modules for all other lineages have the exact opposite configuration, i.e. the genes are down-regulated, their DNA is methylated, and they are in a transcriptionally restrictive chromatin formation. Without any other influences, the system remains in its differentiated cell lineage over time. The corresponding lineage is stable as a cell line would be in reality, i.e. it doesn't spontaneously trans-differentiate or reprogram.

To examine the stability of iPSCs, i.e. when the simulation starts from a state that corresponds to the fully reprogrammed cells where the pluripotency module P has the active configuration, while all other modules are silenced, we can observe a temporal shift of states into states which are closely related to the pluripotency state. Since the states of the model in the epigenetic landscapes from Figure 5.7, as mentioned in the introduction to this Section, have been sorted by their similarity to certain template states, the close relationship is visualized in this figure by their physical proximity to the pluripotency state in the epigenetic landscape. We thus obtain a distribution of states with a high similarity, although not exactly equal, to the pluripotent state of class II iPSCs. This distribution can be observed in iPSCs and ESCs in reality as well and is often referred to as a hyperdynamic plasticity which will quickly be explained in the follwing. IPSCs have a fast changing chromatin structure in general and different methylation states on several loci in the genome (Meshorer et al., 2006). This plasticity can be reflected by the distribution across different states in our model. Since this feature diversifies an ensemble of cells of the same cell type, this effect may also account for the so-called *priming* of iPSCs which allows them to quickly differentiate into a great variety of different cell types upon external signals (Ang et al., 2011). In our simulation, we also observe states that can more easily differentiate than the defined pluripotent state.

Finally, to simulate a reprogramming experiment, we let the model start in the exact state of a differentiated cell lineage with the retroviral pluripotency genes expressed and without epigenetic marks. As shown in Figure 5.7 B, the system quickly evolves from the initial state into transient states that resemble the pluripotent state more and more as time progresses. After a certain time, we can observe how the fully reprogrammed state accumulates, i.e. the system's probability to be in this state increases. This probability can be interpreted as the reprogramming efficiency as it reflects the number of outcomes of a stochastic simulation that would end up in this state. When stochastically simulating a cell population, it could thus be considered as the relative number of cells out of all cells at the beginning that achieve the fully

reprogrammed state. Just as demonstrated earlier by Hanna et al. (2009), this efficiency increases with time (or cell cycles) in our experiment.

The state space of the simulation in Figure 5.3 retraces the timing of reprogramming which is the sequence of states crossed in a simulation from differentiated to pluripotent cells. Our model contains 14 variables, thus the state space has $2^{14} = 16384$ states with a wealth of connections between each other which would be difficult to represent and draw conclusions of it. This is why, we decided to only display states that are reached in the simulation with a probability of at least 0.0001. To our satisfaction the timing of states in our simulation reflects events that also are important in the reprogramming process in reality and are close to those described in literature (Papp and Plath, 2011). Our simulation comprises 500 time steps after which most probabilities only change slightly anymore. In the beginning, i.e. approximately in the first 100 time steps, we can observe how epigenetic marks are slowly removed from the pluripotency module. In the next roughly 150 time steps, expression of differentiation related genes is strongly down-regulated while the pluripotency related genes are still not expressed. At this stage, we observe the accumulation of a dead-end attractor state in which none of the modules shows an expression. This state thus represents a clear roadblock to reprogramming and will be discussed further below. Fortunately with a much higher possibility, the following phase is represented by the increase in probability of states that can be classified class I iPS cells (Niwa, 2007b; Mikkelsen et al., 2008) which show expression of the endogenous pluripotency genes while the retroviral transcripts are still active as well. Subsequently the latter are continuously silenced by epigenetic marks thus reaching the class II iPS cell (or fully reprogrammed) state. Our subspace of the state space in Figure 5.3 contains 146 states and 2473 edges which is approximately half of the states that could theoretically be reached from the start state. Since the overall number of possible transitions between states is incredibly high, every state transition has a very low probability and every path crossing several states has an even lower probability. Hence, the most likely straight path from the differentiation related start state to the fully reprogrammed one only has a probability of $9.3 \cdot 10^{-12}$ and consists of 7 state transitions (shown as a thick line in Figure 5.3). At a closer look, it becomes clear, that in the state space there are a lot of different roads leading to Rome, Rome being the reprogrammed state and that there are faster and less fast roads and transitions that are more or less important, i.e. more or less probable. There is, however, one state transition that cannot be neglected and that all paths will have to cross at one moment. It is like an enormous crossing where all states arrive and from where paths lead in a lot of different directions but where everyone has to pass and there is no shortcut. This transition is essential for reprogramming and it consists of the early activation of the pluripotency module (or maybe in reality some of its actors such as POU5F1

which is strongly up-regulated 96h after transduction in the *OCT4* and the reprogramming (*3TF* and *4TF*) conditions from Chapter 4) as can be seen in Table A.1) in order to down-regulate expression of the differentiated genes upon the removal of their epigenetic marks.

As stated before during the analysis of the landscapes, the state space as well shows that after 500 time steps when the end of the process approaches, the probability that cells are reprogrammed has an increasing tendency as was found by Hanna et al. (2009), who showed that in a reprogramming system with inducible vectors, all cells have the potential to reprogram if they are have enough time (corresponding to a certain number of cell cycles).

We also found that in simulations where all modules are demethylated in the early phase, the pluripotency genes are activated much faster compared to the rest which strengthens the hypothesis that epigenetics, and maybe especially DNA methylation marks, really are the efficiency- and time-limiting step in reprogramming.

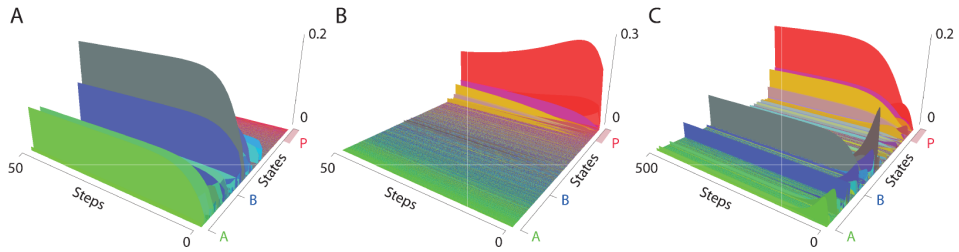

Figure 5.6: Simulations of the model with initial distributions (taken from Flöttmann, Scharp, and Klipp (2012))

A Simulation starting from a distribution of states around the differentiation related state (B) without retroviral reprogramming factors. The states re-distribute around the two differentiation related states (A) and (B) and accumulate in an undefined state that will be explained later. **B** When starting from a distribution around the pluripotent state, the probabilities are re-distributed into states around and thus similar to the pluripotent state confirming the hyperdynamic plasticity (see text) of this state. **C** With active reprogramming factors, a simulation starting from a distribution around the differentiated state (B) yields a reprogramming experiment with the pluripotent state accumulating over time beside other states that also reach a non-negligible probability in the time course.

As mentioned earlier, in the introduction to this Section, if we are dealing with cell populations, stochasticity in cellular processes will always lead to a diversity of cells even inside the same lineage. Although the main expression program will be very similar, genes that don't play a strong role might or might not be expressed, epigenetic marks might or might not or might partially be set inside promoter regions and so on. This is why, we included a certain stochasticity inside cell lineages by generating distributions around a sharp state. The generation of these starting distributions is explained

in more detail in the introduction to this Section. Interestingly, just as for the sharp state simulation, when we start our simulations from these distributions for the differentiated cell lineages without retroviral genes we can observe that a distribution around the differentiated state is maintained accounting for stable cell lineages even in this stochastic case. The same holds for the simulation from a distribution around the pluripotent state P. The system reaches a similar hyperdynamic plasticity distribution as mentioned above in the sharp state case. It is clear that we have to assure that the retroviral genes are silenced when starting out of distributions around the differentiated states since their expression ultimately results in reprogramming (as can be seen in Figure 5.6).

To test our model's behaviour upon perturbations, we are going to analyze parameter variations and structural modifications in the following. We will take a closer look at the strength of epigenetic modifications, i.e. we are going to attribute higher or lower probabilities to the functions including DNA methylation and chromatin formation. Moreover we will examine modificated models in which spontaneous stochastic methylation, demethylation, chromatin formation, no methylation at all or a stronger crosstalk between methylation and chromatin formation occur as structural changes. We will qualitatively evaluate their effects on the reprogramming process, especially its efficiency.

5.5 Parameter Variations of the Model

Considering the strength of epigenetic modifications, we only changed parameters in our main model, i.e. attributed different probabilities to Boolean update functions corresponding to methylation or heterochromatin formation. We thus effected 4 new simulations with higher or lower probabilities than in our main model for both mechanisms respectively. The qualitative effect on the reprogramming efficiency can be seen in Figure 5.8.

As it turns out when looking at an increase in the probability for permissive chromatin formation (called *faster chromatin changes* in Figure 5.8), our main model was apparently already close to maximal saturation after 2000 time steps since the mentioned modification doesn't change the efficiency. Such a modification could experimentally be reached by application of valproic acid (VPA), which is a HDAC1 inhibitor. HDAC1 in general favours heterochromatin formation. Its inhbition is thus comparable with the above mentioned modification. In a different way, speeding up methylation dynamics by higher probabilities of demethylation yields a considerably earlier time of half-maximal efficiency with a slight decrease of maximal efficiency.

On the other hand, we can observe that an increase in the probability for

heterochromatin formation and DNA methylation (called *slower chromatin changes* and *slower DNA methylation*) slows down the reprogramming process considerably. Although the time of half-maximal efficiency is strongly delayed in the latter two cases, it appears that the maximal efficiency that can be reached is very similar to the one of the main model if we artificially stretch the lines past 2000 time steps. Since faster and slower epigenetic changes could also be identified with accelerated or decelerated cell cycle times, this fact strongly recalls the findings by Hanna et al. (2009) (see also Table 5.3).

5.6 Structural Modifications of the Model

We will now focus on the structural modifications mentioned further above and their effects on the system's efficiency. Table 5.3 summarizes the results and as to which degree these modifications reflect known experiments from literature.

5.6.1 Spontaneous Methylation

First, we are considering spontaneous methylation which we introduced by attributing a probability for the methylation states of modules to be activated independently of all other factors except *dnmt*. This introduces a certain measure of stochasticity into the DNMT3A/B dependent *de novo* methylation process which, in fact, is still only poorly understood.

The effects of this modifications are a drastic decrease of reprogramming efficiency approximately by the factor 10 although half-maximal saturation occurs slightly faster than in the main model. As already mentioned above in the explanation of the reprogramming simulation of our main model, in this modification there is a strong accumulation of an undefined state in which all the modules are silenced except for the exogenous one. Further discussion of this undefined roadblock state will be done below further below after having discussed the other structural modifications.

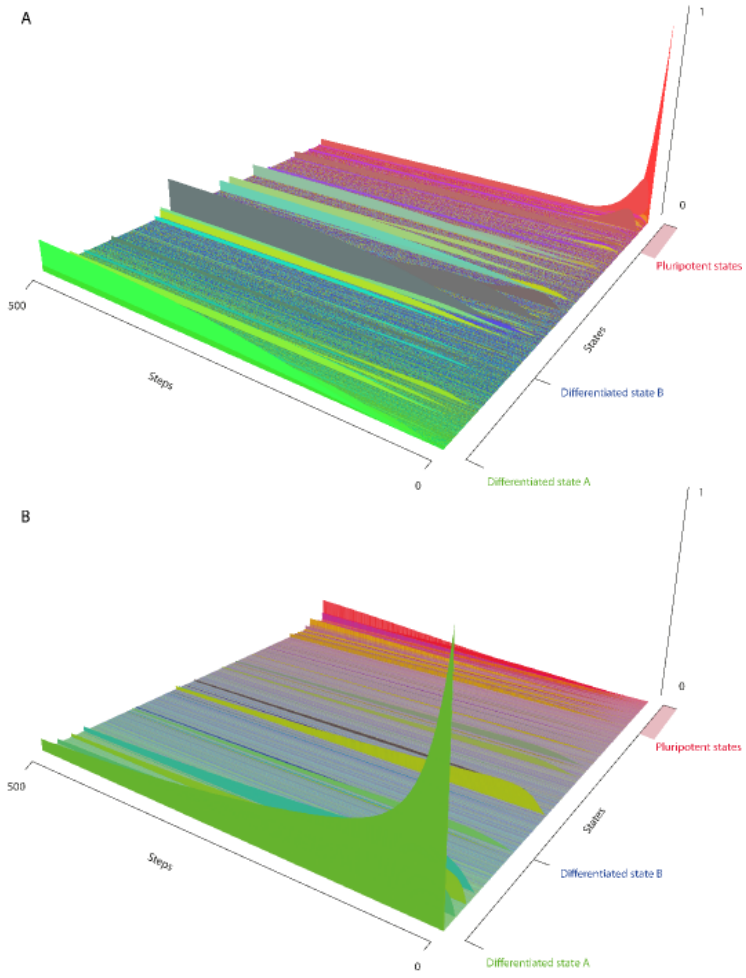

Figure 5.7: Epigenetic Landscapes of Reprogramming and Differentiation (taken from Flöttmann, Scharp, and Klipp (2012))

Shown are 3-dimensional graphs of time courses simulating a differentiation (A) and a reprogramming experiment (B). On the x-axis, the 16384 states of the model are sorted according to their similarity Σ_i^{123} (see Section 2.3.5) to the 3 states that are pointed out on the x-axis. The y-axis is the time line and shows the 500 time steps of the simulation. The z-axis shows the probability of the network to be in a state. The states are furthermore color-coded with colors transitioning from green (differentiated state A) over blue (differentiated state B) to red (pluripotent state). **A** A simulation starting from a unique state, the pluripotent state (see Table 5.2) but with the differentiation module A activated which is needed in order to give a direction to the differentiation is done by signaling pathways in vivo. One can see that the network quickly leaves the pluripotent state and transitions into a few states, the highest probabilities being reached by the undefined state (see text) and the differentiated state A. **B** A simulation starting from the differentiated state A (see Table 5.2) but with the retroviral reprogramming genes being active. We can observe a slow transition from the differentiated state A into states that more and more resemble the pluripotent state and a final progressive accumulation of the latter.

5.6.2 Spontaneous Heterochromatin Formation

When DNA methylation is poorly understood, the wealth of chromatin modifications and their exact effects on DNA packaging and transcriptional regulation add an even more complex layer to the epigenetic jungle. As for the spontaneous DNA methylation, we introduced stochasticity by spontaneous heterochromatin formation, thus partly depriving the process of any external regulation.

The effect of this modification is even stronger than for the de-regulation of DNA methylation and decreases the overall reprogramming efficiency by the factor 40. However, the early first 50 time steps of the process see a faster accumulation of reprogrammed states than the original model. Since heterochromatin formation now occurs spontaneously, there is a considerable probability that differentiation related genes will be shut down via the epigenetic silencing, which leads the system to quickly attain this crucial step in reprogramming. This could explain the faster reprogramming at the beginning. However, the blocking effects of the de-regulation will become apparent rapidly afterwards. Furthermore, we again observe a strong accumulation of the undefined state.

5.6.3 Spontaneous Demethylation

While a lot of epigenetic modifications have been studied for many years, researchers have only recently gained more interest in DNA demethylation processes with some new findings even suggesting that there are enzymatic processes catalysing active demethylation in contrast to cell cycle dependent passive demethylation by decreased DNMT1 activity (see Table 5.1) (Bhutani et al., 2011). We tried to reflect this uncertainty in knowledge on passive or active processes by including a spontaneous demethylation feature as one model variant.

The latter variant reaches the highest reprogramming efficiency after 500 time steps among all model modifications with an efficiency which is approximately 3-4 times lower than in the main model. Moreover, in contrast to the other variants, the spontaneous demethylation model also shows a similar behavior as the original model which is especially characterized by a fast initial decay of the differentiated state followed by a slower decreasing phase.

5.6.4 Stronger Interaction Between Methylation and Heterochromatin

As a next model variant, we examined how a stronger synergistic effect between DNA methylation and chromatin formation (described as well in Table 5.1) influences the reprogramming efficiency. The reprogramming efficiency over time is very similar to the one of spontaneous demethylation described in the Section before this one as are the dynamics of the differentiated state that also show a strong similarity to the original model. Intriguingly, when starting the experiment from cell lineage A, a state similar to the one of differentiated cell lineage B - with the only difference that the pluripotency module is already deprived of its epigenetic marks - is transiently reached with a high probability before it decreases again over time to be transformed steadily into more pluripotency related states. This phenomenon strongly recalls direct biological trans-differentiation of cells during reprogramming which was thought to only work by passing the pluripotent states before the findings of Vierbuchen et al. (2010) who were able to directly convert fibroblasts into functional neurons by defined factors.

5.6.5 No Methylation

In a more drastic modification, we examined the theoretical hypothesis, that DNA methylation has no influence at all on gene expression or chromatin structures. As expected, the model shows a strongly different behavior than before. Leaving out the DNA methylation effects abolishes the ability of the system either to reprogram from a differentiated state with retroviral genes or to differentiate out of a pluripotent state upon signals. We can observe that the start states evolve into a distribution of states that are very closely related just as in the stable cell lineages experiments of the main model. Apparently, without DNA methylation, there is no full silencing of transcriptionally active genes because the crosstalk with chromatin structures and thus heterochromatinization is abolished which is required for complete silencing. Hence, active modules can never be silenced and inactivated gene's expression can never be triggered, even if they are in permissive chromatin structures and their DNA is unmethylated. Therefore, the master regulators of the cell lineage will never change and reprogramming and trans-differentiation are thus impossible.

5.6.6 Polycomb Repressor Complexes (PRCs)

In Table 5.3, we explain the mechanism of Polycomb Repressor Complexes (PRCs). They are in fact epigenetic modifyers that are recruited to the

DNA of differentiation associated and developmental genes upon binding of pluripotency related factors such as OCT4, SOX2 and NANOG. When bound, they modify histone marks in a way as to favour condensed, transcriptionally inactive chromatin. Therefore, to include this PRC mechanism into our model, transformed the equation for heterochromatin formation of the differentiated modules in a way that it positively depends on the expression of the pluriotency module. The model thus created gave a very similar result to the one of our main model (results not shown) suggesting that in our model the mutual transcriptional repression of pluripotency modules and differentiation modules is interchangeable with this PRC mechanism, because they have the same effects. In reality, however, when the system becomes more complex and more tightly regulated, the PRC mechanism might enhance this transcriptional repression and make it more permanent.

5.6.7 Summary of the Model Variants

In Figure 5.8, the effects on the reprogramming efficiency of the different analyzed model variants are shown. Across all variants except for the one without methylation, the reprogramming efficiency generally augments with time although after 2000 time steps it is smaller in nearly every variant than in the main model. The strength of the decrease, however, is very different from variant to variant. As an explanation, it should be noted that all model variants correspond to a more or less strong de-regulation of the main model. More specifically, the epigenetic processes that are tightly regulated in the original model, are rendered more prone to stochasticity which results in the expression of important genes being de-regulated as well. This phenomenon is accompanied with a strong increase of the number of potentially reachable states during the transition. While in the main model a total of 2592 states were reached after 500 time steps in the reprogramming process, in the spontaneous methylation model for example the number increased approximately 4-fold to 10240 states. At the same time the reprogramming efficiency is approximately 10 times lower. It is noteworthy that the 366 pluripotency related states that are reached are the same in both models only differing by their probability after 500 time steps.

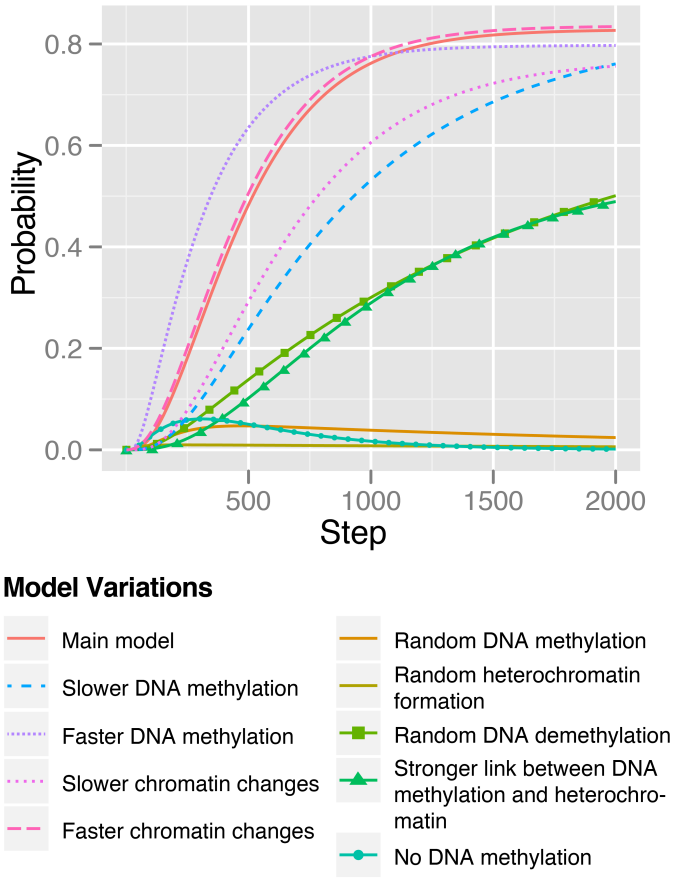

Figure 5.8: Reprogramming efficiencies for the different model variants (taken from Flöttmann, Scharp, and Klipp (2012))

The probability of the network to be in an ensemble of states that are closely related to the pluripotent state (high similarity as defined in Section 2.3.5) is plotted against the simulation time for the main model and its variants that are described in Sections 5.4-5.6.6.

5.7 Summary and Discussion

In this Chapter based on our publication Flöttmann, Scharp, and Klipp (2012), I have outlined the first model of somatic cell reprogramming (to our knowledge) that explicitly includes the retrovirally transduced genes and their regulatory interactions. The model is unique in the way that it introduces the different epigenetic mechanisms that regulate cellular behavior. It is moreover able to qualitatively reproduce experimental results from reprogramming and differentiation experiments. The state space of the PBN together with the dynamic simulation represented as the epigenetic landscape plot provide us with insights into different paths that cells in the process of reprogramming traverse and allow us to identify different milestone phases

during reprogramming. This simulated sequence of events is in accord with the chronological progression reported in reprogramming experiments (see Table 5.3).

In the simulations of our main model, the reprogramming efficiency appears to be very high ($p = 0.8$ after 2000 time steps) compared to real experimental results of reprogramming where it lies below 0.1 most of the time. However, it must be noticed that we are dealing with a highly simplified model leaving out real biological complexity and a wealth of relevant experimental hurdles. The former consists in a much higher number of transcription factors, epigenetic regulators, signaling pathways, micro RNAs to only name a few, while the latter consists in cellular immune responses and low transduction rates for example. The general efficiency shows a similar behavior to experiments done in inducible stem cell systems, which also showed sigmoidal efficiency curves. After long simulation times a steady state with a high amount of reprogrammed cells is reached (as experimentally reached in Hanna et al. (2009)) and these reprogrammed cells consist of broad distribution of pluripotency related states accounting for the hyperdynamic plasticity of pluripotent cells (Niwa, 2007b).

Although our model does not include signaling pathways or other regulating factors controlling cellular differentiation, it is capable of simulating a differentiation experiment that shares many features with the biological process of differentiation. It takes significantly less time than reprogramming but is unspecific and impaired. Including signaling pathways into the model would allow for a more precise activity of the crucial model species and also provide the system with the ability to react to external signaling molecules. To approach biological reality even more, the network model could be extended by further branches of differentiation at the same level and downstream to mimick the progression of differentiation via intermediate cell states with diminishing differentiation potential into various cell lineages. The modular structure of our model simplifies this latter step significantly which is why it could be easily used for future extensions and analyses.

Perturbations and modulations of the model strongly affect reprogramming efficicency and hint at possible points of action for experimental design to improve the process. The strongly negative effect on the efficiency of most modifications indicates the need for tight regulation of the whole transcriptional and epigenetic machinery responsible for cell differentiation and reprogramming. The only two modifications, in which efficiency can be sustained at an adequate level, are those that increase the level of regulation by the genes, namely the random DNA demethylation and the stronger link between DNA methylation and heterochromatin formation.

The reprogramming efficiency could be improved in two modifications of the original model. For the faster change in DNA methylation, the half-maximal

saturation of the sigmoidal efficiency is reached way earlier than in the main model, i.e. reprogrammed cells appear earlier. However, the saturation level at the steady state is slightly lower. This shows that a de-regulation can have a short-term beneficial effect, i.e. in experiments, iPSCs would appear earlier and the process would be accelerated. However, one would have to accept a lower overall number of iPSCs. At the same time, the modification for faster changes in chromatin state nearly have no effect and only very slightly increase the overall efficiency.

Improving our understanding of the detailed mechanisms underlying somatic cell reprogramming is the key to enhancing it and reduce the roadblocks and inconvenient features that still hinder clinical application of iPS cells in the future. The model that we developed in this study might be a good starting point to broaden our knowledge and extend models focusing on one feature such as transcription to multi-feature frameworks including the important epigenetic aspects. It is able to reproduce and explain experimental observations concerning epigenetics and their internal connections as well as those to transcriptional processes while leaving out detailed transcriptional interaction networks and signaling pathways.

Table 5.3: Experimental findings from literature compared to simulation results from our model (taken from Flöttmann, Scharp, and Klipp (2012))

Experimental Finding	**Model validation**
Somatic cells can be reprogrammed to iPSCs upon viral delivery of pluripotency factors with a very low efficiency (Takahashi and Yamanaka, 2006)	Reprogramming experiment of our main model (Figure 5.7 B)
iPSCs can be re-differentiated into various kinds of tissues (all three germ layers) (Takahashi and Yamanaka, 2006)	Differentiation experiment of our main model (Figure 5.7 A)
ESCs have more euchromatin and accumulate high condensed heterochromatin as differentiation progresses (Francastel et al., 2000)	In the differentiation of the pluripotent state, which still consists of a distribution across several different chromatin and methylation configurations, we can observe a transition to more sharply defined states, which mostly include heterochromatin and methylation compositions (Figure 5.7 A)
DNA methylation is essential for chromatin structure during development (Hashimshony et al., 2003)	In models lacking DNA methylation, differentiation as well as reprogramming are abolished and cells will not be able to pass to other states in the state space (Section 5.6.5)
Treatment of partially differentiated ES cells with the DNA demethylating agent 5-azacytidine (5-AzaC) induces de-differentiation (Tsuji-Takayama et al., 2004)	When starting from partly differentiated states in models with spontaneous demethylation mimicking 5-AzaC treatment, we observe de-differentiation and even efficient reprogramming (Section 5.6.3)
Knockdown of DnmtI reactivates retroviral genes (Wernig et al., 2007)	In models mimicking DnmtI knockdown (e.g. spontaneous demethylation in Section 5.6.3 or no methylation in Section 5.6.5 simulation from the iPS state leads to partial reactivation of retroviral genes

Experimental Finding	**Model validation**
Dnmt3a and Dnmt3b are not required for retroviral silencing in the first 10 days of reprogramming (Pannell et al., 2000; Hotta and Ellis, 2008)	In models without dnmt activity we can still observe silencing of retroviral genes (results not explicitly shown)
The histone deacetylase (HDAC) inhibitor valproic acid is capable of enhancing reprogramming efficiency (Huangfu et al., 2008)	In models where the probability for heterochromatin formation is downregulated (mimicking inhibition of HDAC) we observe a slight increase in the reprogramming efficiency (Figure 5.8).
Polycomb Repressor Complexes (PRCs) are recruited to differentiation associated genes upon binding of pluripotency master regulators (OCT4, SOX2, NANOG) and mediate transcriptional repression in mammals through PRC2-induced H3K27 trimethylation inducing recruitment of PRC1 and subsequent H2A ubiquitinylation leading to chromatin condensation (Boyer et al., 2006)	Including the mechanism of PRCs into our model yields very similar results as the main model suggesting that pure transcriptional repression between master regulators of pluripotency and differentiated lineages is exchangeable with the PRC mechanism although the latter may be more permanent due to the epigenetic features (results not explicitly shown).

6 Discussion and Outlook

Summary of Results

In this work, I have approached the issues of somatic cell reprogramming at different stages of the process from various modeling angles in order to find answers to crucial questions posed by the process.

The first question that I wanted to answer was how networks that are active in pluripotent cells can unite the concept of stability of lineage decisions with the necessary plasticity of pluripotent cells in topological features. In order to approach this issue, a big iPSC specific interaction network gained via automated literature mining and expert curation from the Genomatix Pathway System GePS (algorithm described in Frisch et al. (2009)) was analyzed with respect to its 3-node network motifs frequency and compared to randomly generated homogeneous Boolean networks. It was found that motifs accounting for increased dynamic stability according to their structural stability score (SSS described by Prill et al. (2005)) were significantly under-represented in the iPSC network while motifs with decreased stability were significantly over-represented compared to the random networks. I hypothesized that this is due to the requirement of dynamic flexibility of a network that is involved in multi-stable processes that account for cell lineage decision making on the one hand and dynamic plasticity of the pluripotent state on the other hand. In fact, pluripotent cells have to be able to quickly differentiate into different cell lineages upon defined triggers. If the pluripotent steady state was very stable and rigid, very strong perturbations would be necessary to lift it out of its low differentiation potential pit and to push it towards one or the other cell lineage. However, minor triggers such as the presence of certain signaling molecules such as BMP4 or TGFβ are able to change the fate of ESCs (Greber et al., 2008). Therefore, it is possible that the under-representation of highly stable motifs and the accumulation of motifs showing lower stability work together to decrease the stability of the pluripotency associated attractor thereby increasing its dynamic plas-

ticity (as opposed to stability and rigidity) and sensitivity to differentiation triggers.

Pursuing this idea of decreased stability in pluripotency related networks, I suggested that random networks showing the hypothesized dynamic behavior of decreased stability, i.e. smaller than expected basin sizes of the pluripotency associated attractor in the corresponding Boolean state space graph, would show a similar distribution of motifs. In other words I assumed that the decreased stability criterion of the pluripotent state alone would suffice to generate networks with the same topological features as the iPSC literature network. This assumption, however, could not fully be approved. A trend is recognizable in which the ensemble of networks that have a lower than average stability of one attractor has a mean relative frequency of occurrence more similar to the one of the iPSC network than to the random networks. It is thus possible that my stability criterion for the filtering of the random networks could partly approach the motif distribution of the pluripotency network. Nonetheless, it should be said that difficulties arose from the small network size of only 10 nodes for the RBNs and thus the networks filtered for their decreased stability. In fact, in such small networks, different scaling effects (Erdős-Rényi scaling for small networks as discussed in 3.3) for motif frequencies than in bigger networks can possibly occur which is why the interpretation of the results should be treated with care. In order to be able to neglect these scaling effects, the size of the RBNs should be increased. However, this increase is limited by the attractor search for the decreased stability filtering.

In summary, the results for the relationship between attractor stability and network topology in the second part might not be pronounced strongly enough to draw a decisive conclusion. Thus far, the decreased stability cannot be taken for granted as the only criterion responsible for the network structure. Nonetheless, the discovered tendencies bear great potential upon further research. Discovering a direct relationship between stability criteria of attractors of the network and topology could facilitate dynamical analyses in the future.

Such a dynamical analysis can only be carried out in a functional model and has the potential to reveal important characteristics of the network and possible steps of underlying mechanisms. Therefore, the big interaction network employed in the motif discovery part summarized above, was thoroughly treated, i.e. filtered, reduced and curated, to yield a highly confident purely transcriptional interaction network. The filtering and reduction was based upon an enrichment with microarray gene expression profiling data for early reprogramming. In fact, only significantly differentially expressed genes were left in the network together with the master regulators of pluripotency. This prior knowledge network (PKN) was then translated into a Boolean model

and combined with the multiple condition reprogramming data in order to optimize it. This model training yielded interesting new insights into early reprogramming, as SP1 emerges to be one of the most prominent switches of the process at this stage. From my optimization results it appears that an initial down-regulation of SP1 via direct inhibition by retroviral KLF4, induces down-regulations of a wealth of genes including IRS1, EPAS1, HIF1A, FGFR1 and c-MYC in a first layer and FGF2 and possibly endogenous KLF4 in a second layer. However, the SP1 dynamics are complex and need to be analyzed in more detail in the future in order to find out when exactly it has to be active promoting pluripotency related processes and when it has to be down-regulated possibly giving rise to hTERT transcription which is crucial for reprogramming.

Beside a new possible activation pathway for the endogenous pluripotency master regulators that includes complex interactions of retroviral OCT4 and KLF4 and endogenous SP1, IRS1 and STAT3, an interesting result is the lack of the prominent interplay between the endogenous master regulators as postulated by Boyer et al. (2005). In fact, it seems that other processes are more important in early reprogramming with the mutual activation of pluripotency master regulators probably being left for later stages. Interesting results were also found with respect to FGF2 and hypoxia inducible factors regulation that are counter-intuitive with their expression in iPSCs. Moreover, the regulation of CCND1 possibly suggests an early reprogramming G_0/G_1 arrest of the cell cycle in those conditions where retroviral *KLF4* is present.

It is further noteworthy that optimizations with different normalization starting points yielded different results. There were in fact two different possible starting conditions, the pure fibroblast measurement on the one hand and the measurement of fibroblasts transduced with a vector only carrying the GFP gene on the other. It could be found that normalizing against the condition with the transduced GFP generally yielded better optimization results. I hypothesized this to be due to the lack of components for the viral response in the interaction network while a viral transduction will generally trigger this response. Therefore, I recommend for future comparative experiments and theoretical validations or optimizations to use a similar normalization approach.

A minimal Boolean model of early reprogramming was derived by continuously removing species that are poorly fitted and have little downstream influence. The resulting minimal Boolean model was then simulated in our in-browser tool BooleSim (Bock, Scharp, Talnikar, and Klipp, 2013) for the different initial experimental conditions. It could thereby be found what exact steps are necessary in early reprogramming to arrive at the state in which cells are likely to be found after 96 hours of reprogramming. While

POU5F1 and CCND1 are up-regulated in a first step via the action of retroviral OCT4 and retroviral as well as endogenous c-MYC in combination with initially expressed SP1, a down-regulation of the latter by the presence of retroviral KLF4 induces down-regulation of endogenous KLF4, endogenous c-MYC and thus CCND1 and POU5F1.

The counter-intuitive transcriptional profiles of many genes together with the possible cell cycle arrest gives me reason to hypothesize the existence of an intermediate reprogramming state with low transcriptional activity of genes that will need to be transcribed later in reprogrammed iPSCs. Such a tense, intermediate state could possibly be identified with a state in which some re-structuring processes in the cell still need to be achieved before pluripotency related genes can unfold their full transcriptional potential.

Such a re-structuring process was studied in the last step, in which the thus far regarded purely transcriptional interaction networks were extended to include epigenetic processes such as DNA methylation and histone modifications leading to changes in the chromatin structure in order to reflect these processes that are crucial in reprogramming and differentiation during cell lineage decisions. We derived a modular probabilistic Boolean model (PBN) including the retrovirally introduced genes as a module, the endogenous pluripotency master regulators and two master regulators of different cell lineages as well as two DNA modifying species, one accounting for DNA methylating reactions, the other for DNA demethylation processes. The analysis of this model yielded interesting pathways through the Boolean state space as a result of simulated reprogramming and differentiation experiments. Different phases of reprogramming could thereby be unraveled whose chronological progression is in strong accord with experimental findings (summarized in Figure 5.3 and Table 5.3). It seems that the first phase consists in the removal of epigenetic repressive marks of the pluripotency master regulators. Subsequently, the master regulators of the initial cell lineage are down-regulated followed by an up-regulation of the pluripotency master regulators leading to class I iPSCs and after epigenetic silencing of the retroviral genes to class II iPSCs.

Modifying the model structure and parameters to reflect changes in the interplay between the epigenetic processes showed that manipulations of regulations could either enhance but in most cases strongly impair the reprogramming efficiency. Since the modifications of the model always consisted in a de-regulation of the mechanisms involved, this finding underlines the necessity for a tight regulation and partly confirms the structure and dynamical behavior of our main model.

Interestingly, our model shows an intermediate state (which was designated as *undesired state* in Chapter 5) during a reprogramming simulation in which all genes are unexpressed. The similarity to the intermediate state from the

optimization of the Boolean pluripotency model in which all target genes are transiently down-regulated is striking although it might just be an interesting coincidence. In the reprogramming experiment of the PBN model, the transient down-regulation of genes is a means to epigenetically re-structure the different gene modules in order to silence master regulators of cell lineages and thus allow pluripotency markers to be expressed without the transcriptional inhibition by the former. It is possible that the ensemble of genes that were found to be transiently down-regulated in the optimizations in Chapter 4 somehow interfere with epigenetic re-structuring processes and thus have to be down-regulated for during this step. It is well known that epigenetic modificators such as valproic acid (VPA) have the potential to strongly enhance the process (Huangfu et al., 2008). It could thus be interesting to compare transcriptional profiles of reprogramming experiments with and without these small chemical compounds in order to unravel whether this intermediate state would still persist.

Outlook

Following the interpretation and summary of this work's results, it is possible to draw conclusions and carefully predict future experiments and possible enhancements of strategies. It will thus be interesting to compare data of later reprogramming stages with the optimization results in this study in order to decipher the exact order of events and test the hypothesis of the existence of an intermediate state with low transcriptional activity. This could help to understand the surprising transient down-regulation of the majority of genes that need to be active in iPSCs. It could moreover be tested whether this intermediate state is really necessary or whether it could be surpassed. In order to do so, I recommend to keep SP1 constitutively active during the reprogramming process which I believe to prevent down-regulation of the majority of its targets and thus possibly of the intermediate state. In order to do so, the inhibition of SP1 by retroviral KLF4 should be prevented. This could either lead to a strong enhancement of the reprogramming process or to a complete abolishing of the latter. Either way, the result would contain a wealth of predictive power on the nature of underlying mechanisms.

It will be of great interest as well to further our understanding of the epigenetic reprogramming and its exact relationship to transcriptional reprogramming in order to understand the existence of the intermediate state during the reprogramming simulation and whether it can be identified with the one found in the optimization experiments. It is too early to draw conclusions about a possible necessary existence of such an intermediate state for a successful reprogramming. However, its occurrence and its presence in experiments needs testing and could bear great potential.

Concluding Remarks

With all the ongoing controversies, the contradictory experimental results and the great potential that holds the reprogramming process, it is out of the question, that a deeper understanding of the mechanisms and exact series of events is needed in order to enhance and be able to clinically apply iPSC therapy at some point in the distant future. In this work, I have shown that a lot of insights into the detailed processes can be gained when combining experimental knowledge with theoretical mathematical modeling. Instead of relying on the qualitative interpretation of experimental results, the systems biology approach provides us with a much more profound and exact understanding of the underlying mechanisms in dynamic biological systems and becomes a more and more an important and necessary complementary part to purely experimental research.

List of Figures

List of Tables

Bibliography

Adewumi, O., B. Aflatoonian, and L. Ahrlund-Richter (2007, July). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature 25*(7), 803–816.

Adnane, J., Z. Shao, and P. D. Robbins (1999, Jan). Cyclin d1 associates with the tbp-associated factor taf(ii)250 to regulate sp1-mediated transcription. *Oncogene 18*(1), 239–247.

Albert, I., J. Thakar, S. Li, R. Zhang, and R. Albert (2008). Boolean network simulations for life scientists. *Source Code Biol Med 3*, 16.

Ammanamanchi, S., S. J. Kim, L. Z. Sun, and M. G. Brattain (1998, Jun). Induction of transforming growth factor-beta receptor type ii expression in estrogen receptor-positive breast cancer cells through sp1 activation by 5-aza-2'-deoxycytidine. *J Biol Chem 273*(26), 16527–16534.

Ang, Y.-S., A. Gaspar-Maia, I. R. Lemischka, and E. Bernstein (2011, July). Stem cells and reprogramming: breaking the epigenetic barrier? *Trends in pharmacological sciences 32*(7), 394–401.

Artyomov, M., A. Meissner, and A. Chakraborty (2010, May). A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency. *PLoS computational biology 6*(5), e1000785.

Babaie, Y., R. Herwig, B. Greber, T. C. Brink, W. Wruck, D. Groth, H. Lehrach, T. Burdon, and J. Adjaye (2007, Feb). Analysis of oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells 25*(2), 500–510.

Barabasi and Albert (1999, Oct). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Barabási, A.-L. and Z. N. Oltvai (2004, Feb). Network biology: understanding the cell's functional organization. *Nat Rev Genet 5*(2), 101–113.

Bartlett, M. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A. 160*, 268–282.

Berg, J. M., J. L. Tymoczko, and L. Stryer (2002). *Biochemistry* (5th ed.). New York: W H Freeman.

Berger, S. L., T. Kouzarides, R. Shiekhattar, and A. Shilatifard (2009, Apr). An operational definition of epigenetics. *Genes Dev 23*(7), 781–783.

Bhutani, N., D. Burns, and H. Blau (2011, September). DNA Demethylation Dynamics. *Cell 146*(6), 866–872.

Black, S. M., J. M. DeVol, and S. Wedgwood (2008, Jan). Regulation of fibroblast growth factor-2 expression in pulmonary arterial smooth muscle cells involves increased reactive oxygen species generation. *Am J Physiol Cell Physiol 294*(1), C345–C354.

Bland, M. (1995). *An Introduction to Medical Statistics* (3 (2000) ed.). Oxford Medical Publications.

Bock, M., T. Scharp, C. Talnikar, and E. Klipp (2013, Sep). Boolesim: An interactive boolean network simulator. *Bioinformatics*.

Bollobas, B. (1985). *Random Graphs*. Academic Press, New York.

Bourillot, P.-Y., I. Aksoy, V. Schreiber, F. Wianny, H. Schulz, O. Hummel, N. Hubner, and P. Savatier (2009, Aug). Novel stat3 target genes exert distinct roles in the inhibition of mesoderm and endoderm differentiation in cooperation with nanog. *Stem Cells 27*(8), 1760–1771.

Bowman, T., M. A. Broome, D. Sinibaldi, W. Wharton, W. J. Pledger, J. M. Sedivy, R. Irby, T. Yeatman, S. A. Courtneidge, and R. Jove (2001, Jun). Stat3-mediated myc expression is required for src transformation and pdgf-induced mitogenesis. *Proc Natl Acad Sci U S A 98*(13), 7319–7324.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young (2005, September). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell 122*(6), 947–956.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch (2006, May). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature 441*(7091), 349–353.

Brambrink, T., R. Foreman, G. G. Welstead, C. J. Lengner, M. Wernig, H. Suh, and R. Jaenisch (2008, Feb). Sequential expression of pluripotency

markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell 2*(2), 151–159.

Bártová, E., J. Krejcí, A. Harnicarová, G. Galiová, and S. Kozubek (2008, Aug). Histone modifications and nuclear architecture: a review. *J Histochem Cytochem 56*(8), 711–721.

Bäck, T. and F. Hoffmeister (1991). Extended selection mechanisms in genetic algorithms. pp. 92–99. Morgan Kaufmann.

Carbone, M., M. N. Rossi, M. Cavaldesi, A. Notari, P. Amati, and R. Maione (2008, Oct). Poly(adp-ribosyl)ation is implicated in the g0-g1 transition of resting cells. *Oncogene 27*(47), 6083–6092.

Cedar, H. and Y. Bergman (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics 10*(5), 295–304.

Chang, R., R. Shoemaker, and W. Wang (2011, December). Systematic Search for Recipes to Generate Induced Pluripotent Stem Cells. *PLoS Computational Biology 7*(12), e1002300.

Chavez, L., A. Bais, M. Vingron, H. Lehrach, J. Adjaye, and R. Herwig (2009). In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics 10*(1), 314.

Chen, H.-F., H.-C. Kuo, W. Chen, F.-C. Wu, Y.-S. Yang, and H.-N. Ho (2009, Jan). A reduced oxygen tension (5embryonic stem cells in the undifferentiated state with short splitting intervals. *Hum Reprod 24*(1), 71–80.

Chen, Y., L. Shi, L. Zhang, R. Li, J. Liang, W. Yu, L. Sun, X. Yang, Y. Wang, Y. Zhang, and Y. Shang (2008, Jun). The molecular mechanism governing the oncogenic potential of sox2 in breast cancer. *J Biol Chem 283*(26), 17969–17978.

Chickarmane, V. and C. Peterson (2008). A Computational Model for Understanding Stem Cell, Trophectoderm and Endoderm Lineage Determination. *PLoS ONE 3*(10), e3478.

Chickarmane, V., C. Troein, and U. Nuber (2006). Transcriptional dynamics of the embryonic stem cell switch. *PLoS computational 2*(9), e123.

Chuang, H.-Y., M. Hofree, and T. Ideker (2010). A decade of systems biology. *Annu Rev Cell Dev Biol 26*, 721–744.

Ciriello, G. and C. Guerra (2008, Mar). A review on models and algorithms for motif discovery in protein-protein interaction networks. *Brief Funct Genomic Proteomic 7*(2), 147–156.

Coma, S., D. N. Amin, A. Shimizu, A. Lasorella, A. Iavarone, and M. Klagsbrun (2010, May). Id2 promotes tumor cell migration and invasion through transcriptional repression of semaphorin 3f. *Cancer Res 70*(9), 3823–3832.

Conover, W. (1980). *Practical Nonparametric Statistics*. John Wiley and Sons, New York.

Covello, K. L., J. Kehler, H. Yu, J. D. Gordan, A. M. Arsham, C.-J. Hu, P. A. Labosky, M. C. Simon, and B. Keith (2006, Mar). Hif-2alpha regulates oct-4: effects of hypoxia on stem cell function, embryonic development, and tumor growth. *Genes Dev 20*(5), 557–570.

Cram, E. J., B. D. Liu, L. F. Bjeldanes, and G. L. Firestone (2001, Jun). Indole-3-carbinol inhibits cdk6 expression in human mcf-7 breast cancer cells by disrupting sp1 transcription factor interactions with a composite element in the cdk6 gene promoter. *J Biol Chem 276*(25), 22332–22340.

Crick, F. (1970, Aug). Central dogma of molecular biology. *Nature 227*(5258), 561–563.

Crosby, J. L. (1973). *Computer Simulation in Genetics*. London: John Wiley & Sons.

Dalton, S. (2013, Apr). Signaling networks in human pluripotent stem cells. *Curr Opin Cell Biol 25*(2), 241–246.

Do, D. V., J. Ueda, D. M. Messerschmidt, C. Lorthongpanich, Y. Zhou, B. Feng, G. Guo, P. J. Lin, M. Z. Hossain, W. Zhang, A. Moh, Q. Wu, P. Robson, H. H. Ng, L. Poellinger, B. B. Knowles, D. Solter, and X.-Y. Fu (2013, Jun). A genetic and developmental pathway from stat3 to the oct4-nanog circuit is essential for maintenance of icm lineages in vivo. *Genes Dev 27*(12), 1378–1390.

Dodd, I. B., M. A. Micheelsen, K. Sneppen, and G. Thon (2007, May). Theoretical Analysis of Epigenetic Cell Memory by Nucleosome Modification. *Cell 129*(4), 813–822.

Drossel, B., T. Mihaljev, and F. Greil (2005, Mar). Number and length of attractors in a critical kauffman model with connectivity one. *Phys Rev Lett 94*(8), 088701.

Elser, M., L. Borsig, P. O. Hassa, S. Erener, S. Messner, T. Valovka, S. Keller, M. Gassmann, and M. O. Hottiger (2008, Feb). Poly(adp-ribose) polymerase 1 promotes tumor cell survival by coactivating hypoxia-inducible factor-1-dependent gene expression. *Mol Cancer Res 6*(2), 282–290.

Eminli, S., J. Utikal, K. Arnold, R. Jaenisch, and K. Hochedlinger (2008, Oct). Reprogramming of neural progenitor cells into induced pluripotent

stem cells in the absence of exogenous sox2 expression. *Stem Cells 26*(10), 2467–2474.

Epsztejn-Litman, S., N. Feldman, M. Abu-Remaileh, Y. Shufaro, A. Gerson, J. Ueda, R. Deplus, F. Fuks, Y. Shinkai, H. Cedar, and Y. Bergman (2008, November). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nature structural & molecular biology 15*(11), 1176–83.

Ezashi, T., P. Das, and R. M. Roberts (2005, Mar). Low o2 tensions and the prevention of differentiation of hes cells. *Proc Natl Acad Sci U S A 102*(13), 4783–4788.

Fauré, A., A. Naldi, C. Chaouiya, and D. Thieffry (2006, Jul). Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics 22*(14), e124–e131.

Feldman, N., A. Gerson, J. Fang, E. Li, Y. Zhang, Y. Shinkai, H. Cedar, and Y. Bergman (2006, Feb). G9a-mediated irreversible epigenetic inactivation of oct-3/4 during early embryogenesis. *Nat Cell Biol 8*(2), 188–194.

Firestone, G. L. and L. F. Bjeldanes (2003, Jul). Indole-3-carbinol and 3-3'-diindolylmethane antiproliferative signaling pathways control cell-cycle gene transcription in human breast cancer cells by regulating promoter-sp1 transcription factor interactions. *J Nutr 133*(7 Suppl), 2448S–2455S.

Flöttmann, M., T. Scharp, and E. Klipp (2012). A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front Physiol 3*, 216.

Forristal, C. E., K. L. Wright, N. A. Hanley, R. O. C. Oreffo, and F. D. Houghton (2010, Jan). Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. *Reproduction 139*(1), 85–97.

Foshay, K. M. and G. I. Gallicano (2008, Apr). Regulation of sox2 by stat3 initiates commitment to the neural precursor cell fate. *Stem Cells Dev 17*(2), 269–278.

Fraga, M. F., M. Berdasco, E. Ballestar, S. Ropero, P. Lopez-Nieva, L. Lopez-Serra, J. I. Martín-Subero, M. J. Calasanz, I. L. de Silanes, F. Setien, S. Casado, A. F. Fernandez, R. Siebert, S. Stifani, and M. Esteller (2008, Jun). Epigenetic inactivation of the groucho homologue gene tle1 in hematologic malignancies. *Cancer Res 68*(11), 4116–4122.

Francastel, C., D. Schuebeler, D. I. Martin, and M. Groudine (2000, Nov). Nuclear compartmentalization and gene activity. *Nat Rev Mol Cell Biol 1*(2), 137–143.

Fraser, A. and D. Burnell (1970). *Computer Models in Genetics*. New York: McGraw Hill.

Frederick, J. P., N. T. Liberati, D. S. Waddell, Y. Shi, and X.-F. Wang (2004, Mar). Transforming growth factor beta-mediated transcriptional repression of c-myc is dependent on direct binding of smad3 to a novel repressive smad binding element. *Mol Cell Biol 24*(6), 2546–2559.

Frisch, M., B. Klocke, M. Haltmeier, and K. Frech (2009, Jul). Litinspector: literature and signal transduction pathway mining in pubmed abstracts. *Nucleic Acids Res 37*(Web Server issue), W135–W140.

Fuks, F., W. a. Burgers, a. Brehm, L. Hughes-Davies, and T. Kouzarides (2000, January). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nature genetics 24*(1), 88–91.

Gao, F., S. W. Kwon, Y. Zhao, and Y. Jin (2009, Aug). Parp1 poly(adp-ribosyl)ates sox2 to control sox2 protein levels and fgf4 expression during embryonic stem cell differentiation. *J Biol Chem 284*(33), 22263–22273.

Garg, A., K. Mohanram, A. Di Cara, G. De Micheli, and I. Xenarios (2009, June). Modeling stochasticity and robustness in gene regulatory networks. *Bioinformatics (Oxford, England) 25*(12), i101–9.

Geltinger, C., K. Hörtnagel, and A. Polack (1996). Tata box and sp1 sites mediate the activation of c-myc promoter p1 by immunoglobulin kappa enhancers. *Gene Expr 6*(2), 113–127.

Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P.-Z. Wen, H. Brewer, K. Weitz, D. G. Camp, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, and A. J. Lusis (2011, Jun). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet 7*(6), e1001393.

Greber, B., H. Lehrach, and J. Adjaye (2007a, Feb). Fibroblast growth factor 2 modulates transforming growth factor beta signaling in mouse embryonic fibroblasts and human escs (hescs) to support hesc self-renewal. *Stem Cells 25*(2), 455–464.

Greber, B., H. Lehrach, and J. Adjaye (2007b). Silencing of core transcription factors in human ec cells highlights the importance of autocrine fgf signaling for self-renewal. *BMC Dev Biol 7*, 46.

Greber, B., H. Lehrach, and J. Adjaye (2008, Dec). Control of early fate decisions in human es cells by distinct states of tgfbeta pathway activity. *Stem Cells Dev 17*(6), 1065–1077.

Halley, J. D., F. R. Burden, and D. a. Winkler (2009, May). Stem cell decision making and critical-like exploratory networks. *Stem cell research 2*(3), 165–77.

Han, J., P. Yuan, H. Yang, J. Zhang, B. S. Soh, P. Li, S. L. Lim, S. Cao, J. Tay, Y. L. Orlov, T. Lufkin, H.-H. Ng, W.-L. Tam, and B. Lim (2010, Feb). Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature 463*(7284), 1096–1100.

Hanna, J., K. Saha, B. Pando, J. van Zon, C. J. Lengner, M. P. Creyghton, A. van Oudenaarden, and R. Jaenisch (2009, November). Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature advance on*.

Hanna, J., M. Wernig, S. Markoulaki, C.-W. Sun, A. Meissner, J. P. Cassady, C. Beard, T. Brambrink, L.-C. Wu, T. M. Townes, and R. Jaenisch (2007, December). Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science (New York, N.Y.) 318*(5858), 1920–3.

Harvey, I. and T. Bossomaier (1997). *Time Out of Joint: Attractors in Asynchronous Boolean Networks.*, Volume Proceedings of the Fourth European Conference on Artificial Life. MIT Press.

Hashimshony, T., J. Zhang, I. Keshet, M. Bustin, and H. Cedar (2003, Jun). The role of dna methylation in setting up chromatin structure during development. *Nat Genet 34*(2), 187–192.

Hawkins, R. D., G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, and S. Klugman (2010, May). Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. *Cell Stem Cell 6*(5), 479–491.

Hay, D. C., L. Sutherland, J. Clark, and T. Burdon (2004). Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells 22*(2), 225–235.

Hermeking, H., C. Rago, M. Schuhmacher, Q. Li, J. F. Barrett, A. J. Obaya, B. C. O'Connell, M. K. Mateyak, W. Tam, F. Kohlhuber, C. V. Dang, J. M. Sedivy, D. Eick, B. Vogelstein, and K. W. Kinzler (2000, Feb). Identification of cdk4 as a target of c-myc. *Proc Natl Acad Sci U S A 97*(5), 2229–2234.

Hochedlinger, K., Y. Yamada, C. Beard, and R. Jaenisch (2005, May). Ectopic expression of oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell 121*(3), 465–477.

Holliday, R. (1990). Mechanisms for the control of gene activity during development. *Biol Rev Camb Philos Soc 65 (4)*, 431–71.

Hong, H., K. Takahashi, T. Ichisaka, T. Aoi, O. Kanagawa, M. Nakagawa, K. Okita, and S. Yamanaka (2009, August). Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. *Nature 460*(7259), 1132–5.

hong Xu, X. and Z. Zhong (2013, Jun). Disease modeling and drug screening for neurological diseases using human induced pluripotent stem cells. *Acta Pharmacol Sin 34*(6), 755–764.

Hotta, A. and J. Ellis (2008). Retroviral vector silencing during iPS cell induction: An epigenetic beacon that signals distinct pluripotent states. *Journal of Cellular Biochemistry 105*(4), 940–948.

Huang, W., S. Zhao, S. Ammanamanchi, M. Brattain, K. Venkatasubbarao, and J. W. Freeman (2005, Mar). Trichostatin a induces transforming growth factor beta type ii receptor promoter activity and acetylation of sp1 by recruitment of pcaf/p300 to a sp1.nf-y complex. *J Biol Chem 280*(11), 10047–10054.

Huangfu, D., R. Maehr, W. Guo, A. Eijkelenboom, M. Snitow, A. E. Chen, and D. a. Melton (2008, July). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nature biotechnology 26*(7), 795–7.

Huesca, M., L. S. Lock, A. A. Khine, S. Viau, R. Peralta, I. H. Cukier, H. Jin, R. A. Al-Qawasmeh, Y. Lee, J. Wright, and A. Young (2009, Sep). A novel small molecule with potent anticancer activity inhibits cell growth by modulating intracellular labile zinc homeostasis. *Mol Cancer Ther 8*(9), 2586–2596.

Ichida, J. K., J. Blanchard, K. Lam, E. Y. Son, J. E. Chung, D. Egli, K. M. Loh, A. C. Carter, F. P. Di Giorgio, K. Koszka, D. Huangfu, H. Akutsu, D. R. Liu, L. L. Rubin, and K. Eggan (2009, November). A Small-Molecule Inhibitor of Tgf-[beta] Signaling Replaces Sox2 in Reprogramming by Inducing Nanog. *Cell Stem Cell 5*(5), 491–503.

Ingram, P. J., M. P. H. Stumpf, and J. Stark (2006). Network motifs: structure does not determine function. *BMC Genomics 7*, 108.

Itzkovitz, S. and U. Alon (2005, Feb). Subgraphs and network motifs in geometric networks. *Phys Rev E Stat Nonlin Soft Matter Phys 71*(2 Pt 2), 026117.

Itzkovitz, S., R. Milo, N. Kashtan, G. Ziv, and U. Alon (2003, Aug). Subgraphs in random networks. *Phys Rev E Stat Nonlin Soft Matter Phys 68*(2 Pt 2), 026127.

Ivanova, N., R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I. R. Lemischka (2006). Dissecting self-renewal in stem cells with RNA interference. *Nature 442*(7102), 533–538.

Jakel, R. J., B. L. Schneider, and C. N. Svendsen (2004, Feb). Using human neural stem cells to model neurological disease. *Nat Rev Genet 5*(2), 136–144.

James, D., A. J. Levine, D. Besser, and A. Hemmati-Brivanlou (2005, Mar). Tgfbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development 132*(6), 1273–1282.

Jennings, R., M. Alsarraj, K. L. Wright, and T. Muñoz-Antonia (2001, Oct). Regulation of the human transforming growth factor beta type ii receptor gene promoter by novel sp1 sites. *Oncogene 20*(47), 6899–6909.

Kalmar, T., C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. M. Arias (2009, Jul). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol 7*(7), e1000149.

Kanai, M., D. Wei, Q. Li, Z. Jia, J. Ajani, X. Le, J. Yao, and K. Xie (2006, Nov). Loss of krüppel-like factor 4 expression contributes to sp1 overexpression and human gastric cancer development and progression. *Clin Cancer Res 12*(21), 6395–6402.

Kashtan, N., S. Itzkovitz, R. Milo, and U. Alon (2004, Jul). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics 20*(11), 1746–1758.

Kauffman, S. (2004, October). A proposal for using the ensemble approach to understand genetic regulatory networks. *Journal of theoretical biology 230*(4), 581–590.

Kauffman, S. A. (1969, Mar). Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol 22*(3), 437–467.

Kaufman, M. H., E. J. Robertson, A. H. Handyside, and M. J. Evans (1983, Feb). Establishment of pluripotential cell lines from haploid mouse embryos. *J Embryol Exp Morphol 73*, 249–261.

Kaufman, V., T. Mihaljev, and B. Drossel (2005, Oct). Scaling in critical random boolean networks. *Phys Rev E Stat Nonlin Soft Matter Phys 72*(4 Pt 2), 046124.

Kawamura, T., J. Suzuki, Y. V. Wang, S. Menendez, L. B. Morera, A. Raya, G. M. Wahl, and J. C. I. Belmonte (2009, August). Linking the p53 tumour

suppressor pathway to somatic cell reprogramming. *Nature 460*(7259), 1140–4.

Kim, J., J. Chu, X. Shen, J. Wang, and S. H. Orkin (2008, Mar). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell 132*(6), 1049–1061.

Kim, S.-Y., J. W. Kang, X. Song, B. K. Kim, Y. D. Yoo, Y. T. Kwon, and Y. J. Lee (2013, Apr). Role of the il-6-jak1-stat3-oct-4 pathway in the conversion of non-stem cancer cells into cancer stem-like cells. *Cell Signal 25*(4), 961–969.

Kim, S. Y. and J.-W. Park (2010, Dec). Modulation of hypoxia-inducible factor-1α expression by mitochondrial nadp+-dependent isocitrate dehydrogenase. *Biochimie 92*(12), 1908–1913.

Kitano, H. (2002, Mar). Systems biology: a brief overview. *Science 295*(5560), 1662–1664.

Kitazawa, S., R. Kitazawa, and S. Maeda (1999, Oct). Transcriptional regulation of rat cyclin d1 gene by cpg methylation status in promoter region. *J Biol Chem 274*(40), 28787–28793.

Kiuchi, N., K. Nakajima, M. Ichiba, T. Fukada, M. Narimatsu, K. Mizuno, M. Hibi, and T. Hirano (1999, Jan). Stat3 is required for the gp130-mediated full activation of the c-myc gene. *J Exp Med 189*(1), 63–73.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari 4*, 83–91.

Korgaonkar, S. N., X. Feng, M. D. Ross, T. chi Lu, V. D'Agati, R. Iyengar, P. E. Klotman, and J. C. He (2008, May). Hiv-1 upregulates vegf in podocytes. *J Am Soc Nephrol 19*(5), 877–883.

Koshikawa, N., J.-I. Hayashi, A. Nakagawara, and K. Takenaga (2009, Nov). Reactive oxygen species-generating mitochondrial dna mutation up-regulates hypoxia-inducible factor-1alpha gene transcription via phosphatidylinositol 3-kinase-akt/protein kinase c/histone deacetylase pathway. *J Biol Chem 284*(48), 33185–33194.

Kowanetz, M., U. Valcourt, R. Bergström, C.-H. Heldin, and A. Moustakas (2004, May). Id2 and id3 define the potency of cell proliferation and differentiation responses to transforming growth factor beta and bone morphogenetic protein. *Mol Cell Biol 24*(10), 4241–4254.

Krause, F., M. Schulz, B. Ripkens, M. Flöttmann, M. Krantz, E. Klipp, and T. Handorf (2013, Apr). Biographer: web-based editing and rendering of sbgn compliant biochemical networks. *Bioinformatics*.

Kubo, A., K. Shinozaki, J. M. Shannon, V. Kouskoff, M. Kennedy, S. Woo, H. J. Fehling, and G. Keller (2004, Apr). Development of definitive endoderm from embryonic stem cells in culture. *Development 131*(7), 1651–1662.

Kurabayashi, M., S. Dutta, and L. Kedes (1994, Dec). Serum-inducible factors binding to an activating transcription factor motif regulate transcription of the id2a promoter during myogenic differentiation. *J Biol Chem 269*(49), 31162–31170.

Kyo, S., M. Takakura, T. Taira, T. Kanaya, H. Itoh, M. Yutsudo, H. Ariga, and M. Inoue (2000, Feb). Sp1 cooperates with c-myc to activate transcription of the human telomerase reverse transcriptase gene (htert). *Nucleic Acids Res 28*(3), 669–677.

Laniel, M.-A., G. G. Poirier, and S. L. Guérin (2004, Jul). A conserved initiator element on the mammalian poly(adp-ribose) polymerase-1 promoters, in combination with flanking core elements, is necessary to obtain high transcriptional activity. *Biochim Biophys Acta 1679*(1), 37–46.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring, and C.-L. Wei (2010, March). Dynamic changes in the human methylome during differentiation. *Genome Research 20*(3), 320–331.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. Springer.

Levine, M. and E. H. Davidson (2005, Apr). Gene regulatory networks for development. *Proc Natl Acad Sci U S A 102*(14), 4936–4942.

Li, E., T. H. Bestor, and R. Jaenisch (1992, Jun). Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell 69*(6), 915–926.

Li, R., J. Liang, S. Ni, T. Zhou, X. Qing, H. Li, W. He, J. Chen, F. Li, Q. Zhuang, B. Qin, J. Xu, W. Li, J. Yang, Y. Gan, D. Qin, S. Feng, H. Song, D. Yang, B. Zhang, L. Zeng, L. Lai, M. A. Esteban, and D. Pei (2010, Jul). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell 7*(1), 51–63.

Liang, J., M. Wan, Y. Zhang, P. Gu, H. Xin, S. Y. Jung, J. Qin, J. Wong, A. J. Cooney, D. Liu, and Z. Songyang (2008, Jun). Nanog and oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat Cell Biol 10*(6), 731–739.

Lindvall, O. and Z. Kokaia (2006, Jun). Stem cells for the treatment of neurological disorders. *Nature 441*(7097), 1094–1096.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker (2009, November). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature 462*(7271), 315–322.

Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. a. Thomson, R. M. Evans, and J. R. Ecker (2011, February). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*.

Liu, X., J. Huang, T. Chen, Y. Wang, S. Xin, J. Li, G. Pei, and J. Kang (2008, Dec). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Res 18*(12), 1177–1189.

Lo, B. and L. Parham (2009, May). Ethical issues in stem cell research. *Endocr Rev 30*(3), 204–213.

Loh, Y.-H., Q. Wu, J.-L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K.-Y. Wong, K. W. Sung, C. W. H. Lee, X.-D. Zhao, K.-P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C.-L. Wei, Y. Ruan, B. Lim, and H.-H. Ng (2006, April). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet 38*(4), 431–440.

Löfstedt, T., A. Jögi, M. Sigvardsson, K. Gradin, L. Poellinger, S. Påhlman, and H. Axelson (2004, Sep). Induction of id2 expression by hypoxia-inducible factor-1: a role in dedifferentiation of hypoxic neuroblastoma cells. *J Biol Chem 279*(38), 39223–39231.

Macarthur, B. D., A. Ma'ayan, and I. R. Lemischka (2009, Oct). Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol 10*(10), 672–681.

MacArthur, B. D., C. P. Please, and R. O. C. Oreffo (2008). Stochasticity and the Molecular Mechanisms of Induced Pluripotency. *PLoS ONE 3*(8), e3086.

Macía, J., S. Widder, and R. Solé (2009, November). Why are cellular switches Boolean? General conditions for multistable genetic circuits. *Journal of Theoretical Biology 261*(1), 126–135.

Mah, N., Y. Wang, M.-C. Liao, A. Prigione, J. Jozefczuk, B. Lichtner, K. Wolfrum, M. Haltmeier, M. Flöttmann, M. Schaefer, A. Hahn, R. Mrowka, E. Klipp, M. a. Andrade-Navarro, and J. Adjaye (2011, January). Molecular Insights into Reprogramming-Initiation Events Mediated by the OSKM Gene Regulatory Network. *PloS one 6*(8), e24351.

Mahatan, C. S., K. H. Kaestner, D. E. Geiman, and V. W. Yang (1999, Dec). Characterization of the structure and regulation of the murine gene encoding gut-enriched krüppel-like factor (krüppel-like factor 4). *Nucleic Acids Res 27*(23), 4562–4569.

Majello, B., P. D. Luca, G. Suske, and L. Lania (1995, May). Differential transcriptional regulation of c-myc promoter through the same dna binding sites targeted by sp1-like proteins. *Oncogene 10*(9), 1841–1848.

Mann, H. and D. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of mathematical Statistics 18*, 50–60.

Marin, M., A. Karis, P. Visser, F. Grosveld, and S. Philipsen (1997, May). Transcription factor sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell 89*(4), 619–628.

Marión, R. M., K. Strati, H. Li, M. Murga, R. Blanco, S. Ortega, O. Fernandez-Capetillo, M. Serrano, and M. A. Blasco (2009, August). A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature 460*(7259), 1149–53.

Marzec, M., X. Liu, W. Wong, Y. Yang, T. Pasha, K. Kantekure, P. Zhang, A. Woetmann, M. Cheng, N. Odum, and M. A. Wasik (2011, Mar). Oncogenic kinase npm/alk induces expression of hif1α mrna. *Oncogene 30*(11), 1372–1378.

Masui, S., Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. H. Ko, and H. Niwa (2007, Jun). Pluripotency governed by sox2 via regulation of oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol 9*(6), 625–635.

Matsuura, I., N. G. Denissova, G. Wang, D. He, J. Long, and F. Liu (2004, Jul). Cyclin-dependent kinases regulate the antiproliferative function of smads. *Nature 430*(6996), 226–231.

Meng, G., S. Liu, X. Li, R. Krawetz, and D. E. Rancourt (2010, Apr). Extracellular matrix isolated from foreskin fibroblasts supports long-term xeno-free human embryonic stem cell culture. *Stem Cells Dev 19*(4), 547–556.

Meshorer, E., D. Yellajoshula, E. George, P. J. Scambler, D. T. Brown, and T. Misteli (2006, January). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell 10*(1), 105–16.

Mikkelsen, T. S., J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner (2008, July). Dissecting direct reprogramming through integrative genomic analysis. *Nature 454*(7200), 49–55.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon (2002, Oct). Network motifs: simple building blocks of complex networks. *Science 298*(5594), 824–827.

Mitsui, K., Y. Tokuzawa, H. Itoh, K. Segawa, M. Murakami, K. Takahashi, M. Maruyama, M. Maeda, and S. Yamanaka (2003, May). The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and es cells. *Cell 113*(5), 631–642.

Monk, M., R. L. Adams, and A. Rinaldi (1991, May). Decrease in dna methylase activity during preimplantation development in the mouse. *Development 112*(1), 189–192.

Moreno-Manzano, V., F. J. Rodríguez-Jiménez, J. L. Aceña-Bonilla, S. Fustero-Lardíes, S. Erceg, J. Dopazo, D. Montaner, M. Stojkovic, and J. M. Sánchez-Puelles (2010, Jan). Fm19g11, a new hypoxia-inducible factor (hif) modulator, affects stem cell differentiation status. *J Biol Chem 285*(2), 1333–1342.

Müssel, C., M. Hopfensitz, and H. A. Kestler (2010, April). BoolNet - an R package for generation, reconstruction, and analysis of Boolean networks. *Bioinformatics (Oxford, England)*.

Na, J., M. K. Furue, and P. W. Andrews (2010, Sep). Inhibition of erk1/2 prevents neural and mesendodermal differentiation and promotes human embryonic stem cell self-renewal. *Stem Cell Res 5*(2), 157–169.

Nagata, D., E. Suzuki, H. Nishimatsu, H. Satonaka, A. Goto, M. Omata, and Y. Hirata (2001, Jan). Transcriptional activation of the cyclin d1 gene is mediated by multiple cis-elements, including sp1 sites and a camp-responsive element in vascular endothelial cells. *J Biol Chem 276*(1), 662–669.

Nichols, J., B. Zevnik, K. Anastassiadis, H. Niwa, D. Klewe-Nebenius, I. Chambers, H. Schöler, and A. Smith (1998, Oct). Formation of pluripotent stem cells in the mammalian embryo depends on the pou transcription factor oct4. *Cell 95*(3), 379–391.

Nickenig, G., S. Baudler, C. Müller, C. Werner, N. Werner, H. Welzel, K. Strehlow, and M. Böhm (2002, Jul). Redox-sensitive vascular smooth muscle cell proliferation is mediated by gklf and id3 in vitro and in vivo. *FASEB J 16*(9), 1077–1086.

Nishimura, S., S. Takahashi, T. Kuroha, N. Suwabe, T. Nagasawa, and C. Trainor (2000). A GATA Box in the GATA-1 Gene Hematopoietic Enhancer Is a Critical Element in the Network of GATA Factors and Sites That Regulate This Gene A GATA Box in the GATA-1 Gene Hematopoi-

etic Enhancer Is a Critical Element in the Network of GATA Factors and Sites That Regulate This Gene. *Society*.

Nishino, K., M. Toyoda, M. Yamazaki-Inoue, Y. Fukawatase, E. Chikazawa, H. Sakaguchi, H. Akutsu, and A. Umezawa (2011, May). DNA Methylation Dynamics in Human Induced Pluripotent Stem Cells over Time. *PLoS genetics 7*(5), e1002085.

Niwa, H. (2007a, Feb). How is pluripotency determined and maintained? *Development 134*(4), 635–646.

Niwa, H. (2007b, February). How is pluripotency determined and maintained? *Development 134*(4), 635–46.

Niwa, H., K. Ogawa, D. Shimosato, and K. Adachi (2009, Jul). A parallel circuit of lif signalling pathways maintains pluripotency of mouse es cells. *Nature 460*(7251), 118–122.

Niwa, H., Y. Toyooka, D. Shimosato, D. Strumpf, K. Takahashi, R. Yagi, and J. Rossant (2005a, Dec). Interaction between oct3/4 and cdx2 determines trophectoderm differentiation. *Cell 123*(5), 917–929.

Niwa, H., Y. Toyooka, D. Shimosato, D. Strumpf, K. Takahashi, R. Yagi, and J. Rossant (2005b, December). Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation. *Cell 123*(5), 917–929.

Obaya, A. J., M. K. Mateyak, and J. M. Sedivy (1999, May). Mysterious liaisons: the relationship between c-myc and the cell cycle. *Oncogene 18*(19), 2934–2941.

Obokata, H., T. Wakayama, Y. Sasai, K. Kojima, M. P. Vacanti, H. Niwa, M. Yamato, and C. A. Vacanti (2014, Jan). Stimulus-triggered fate conversion of somatic cells into pluripotency. *Nature 505*(7485), 641–647.

Okamoto, A., W. Jiang, S. J. Kim, E. A. Spillare, G. D. Stoner, I. B. Weinstein, and C. C. Harris (1994, Nov). Overexpression of human cyclin d1 reduces the transforming growth factor beta (tgf-beta) type ii receptor and growth inhibition by tgf-beta 1 in an immortalized human esophageal epithelial cell line. *Proc Natl Acad Sci U S A 91*(24), 11576–11580.

Okita, K., M. Nakagawa, H. Hyenjong, T. Ichisaka, and S. Yamanaka (2008, Nov). Generation of mouse induced pluripotent stem cells without viral vectors. *Science 322*(5903), 949–953.

Okuno, Y., G. Huang, F. Rosenbauer, K. Erica, H. S. Radomska, H. Iwasaki, K. Akashi, F. Moreau-gachelin, Y. Li, P. Zhang, D. G. Tenen, Y. Okuno, G. Huang, F. Rosenbauer, E. K. Evans, H. S. Radomska, H. Iwasaki, K. Akashi, F. Moreau-gachelin, and Y. Li (2005). Potential Autoregulation of Transcription Factor PU . 1 by an Upstream Regulatory Element

Potential Autoregulation of Transcription Factor PU . 1 by an Upstream Regulatory Element. *Society*.

Olkin, I. (1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Chapter Howard Levene: Robust tests for equality of variances, pp. 278–292. Stanford University Press.

Ou, J.-N., J. Torrisani, A. Unterberger, N. Provençal, K. Shikimi, M. Karimi, T. J. Ekström, and M. Szyf (2007, May). Histone deacetylase inhibitor trichostatin a induces global and gene-specific dna demethylation in human cancer cell lines. *Biochem Pharmacol 73*(9), 1297–1307.

Pannell, D., C. S. Osborne, S. Yao, T. Sukonnik, P. Pasceri, A. Karaiskakis, M. Okano, E. Li, H. D. Lipshitz, and J. Ellis (2000, Nov). Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code. *EMBO J 19*(21), 5884–5894.

Panno, M. L., L. Mauro, S. Marsico, D. Bellizzi, P. Rizza, C. Morelli, M. Salerno, F. Giordano, and S. Andò (2006, Feb). Evidence that the mouse insulin receptor substrate-1 belongs to the gene family on which the promoter is activated by estrogen receptor alpha through its interaction with sp1. *J Mol Endocrinol 36*(1), 91–105.

Papp, B. and K. Plath (2011, March). Reprogramming to pluripotency: stepwise resetting of the epigenetic landscape. *Cell research 21*(3), 486–501.

Pascal, E. and R. Tjian (1991, Sep). Different activation domains of sp1 govern formation of multimers and mediate transcriptional synergism. *Genes Dev 5*(9), 1646–1656.

Passier, R., L. W. van Laake, and C. L. Mummery (2008, May). Stem-cell-based therapy and lessons from the heart. *Nature 453*(7193), 322–329.

Periyasamy, S., S. Ammanamanchi, M. P. Tillekeratne, and M. G. Brattain (2000, Sep). Repression of transforming growth factor-beta receptor type i promoter expression by sp1 deficiency. *Oncogene 19*(40), 4660–4667.

Prado-Lopez, S., A. Conesa, A. Armiñán, M. Martínez-Losa, C. Escobedo-Lucea, C. Gandia, S. Tarazona, D. Melguizo, D. Blesa, D. Montaner, S. Sanz-González, P. Sepúlveda, S. Götz, J. E. O'Connor, R. Moreno, J. Dopazo, D. J. Burks, and M. Stojkovic (2010, Mar). Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium. *Stem Cells 28*(3), 407–418.

Prill, R. J., P. A. Iglesias, and A. Levchenko (2005, Nov). Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol 3*(11), e343.

Qu, K. and P. Ortoleva (2008, February). Understanding stem cell differentiation through self-organization theory. *Journal of theoretical biology 250*(4), 606–20.

Radaeva, S., R. Sun, H.-N. Pan, F. Hong, and B. Gao (2004, May). Interleukin 22 (il-22) plays a protective role in t cell-mediated murine hepatitis: Il-22 is a survival factor for hepatocytes via stat3 activation. *Hepatology 39*(5), 1332–1342.

Rais, Y., A. Zviran, S. Geula, O. Gafni, E. Chomsky, S. Viukov, A. A. Mansour, I. Caspi, V. Krupalnik, M. Zerbib, I. Maza, N. Mor, D. Baran, L. Weinberger, D. A. Jaitin, D. Lara-Astiaso, R. Blecher-Gonen, Z. Shipony, Z. Mukamel, T. Hagai, S. Gilad, D. Amann-Zalcenstein, A. Tanay, I. Amit, N. Novershtern, and J. H. Hanna (2013, Sep). Deterministic direct reprogramming of somatic cells to pluripotency. *Nature*.

Ralston, A. and J. Rossant (2005, Aug). Genetic regulation of stem cell origins in the mouse embryo. *Clin Genet 68*(2), 106–112.

Razali, N. M. and Y. B. Wah (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and ande rson-darling tests. *Journal of Statistical Modeling and Analytics 2 No.1*, 21–33.

Rekhtman, N., F. Radparvar, T. Evans, N. Rekhtman, F. Radparvar, T. Evans, and A. I. Skoultchi (1999). GATA-1 : functional antagonism in erythroid cells Direct interaction of hematopoietic transcription factors PU . 1 and GATA-1 : functional antagonism in erythroid cells. *Genes & Development*, 1398–1411.

Riggs, A., V. Russo, and M. RA (1996). *Epigenetic mechanisms of gene regulation*. Plainview, N.Y: Cold Spring Harbor Laboratory Press.

Rodolfa, K. (September 30, 2008). *Inducing pluripotency*. StemBook.

Rodríguez, J. L., J. Sandoval, G. Serviddio, J. Sastre, M. Morante, M.-G. Perrelli, M. L. Martínez-Chantar, J. Viña, J. R. Viña, J. M. Mato, M. A. Avila, L. Franco, G. López-Rodas, and L. Torres (2006, Sep). Id2 leaves the chromatin of the e2f4-p130-controlled c-myc promoter during hepatocyte priming for liver regeneration. *Biochem J 398*(3), 431–437.

Rohani, L., A. A. Johnson, A. Arnold, and A. Stolzing (2013, Nov). The aging signature: a hallmark of ips cells? *Aging Cell*.

Ruiz Acero, G. (2012). *Molecular mechanisms involved in the induction of pluripotency*. Ph. d., Freie Universität Berlin.

Saez-Rodriguez, J., L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt, and P. K. Sorger (2009). Discrete logic modelling as

a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol 5*, 331.

Saez-Rodriguez, J., A. Goldsipe, J. Muhlich, L. G. Alexopoulos, B. Millard, D. A. Lauffenburger, and P. K. Sorger (2008, Mar). Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics 24*(6), 840–847.

Saez-Rodriguez, J., L. Simeoni, J. A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U.-U. Haus, R. Weismantel, E. D. Gilles, S. Klamt, and B. Schraven (2007, Aug). A logical model provides insights into t cell receptor signaling. *PLoS Comput Biol 3*(8), e163.

Samavarchi-Tehrani, P., A. Golipour, L. David, H.-K. Sung, T. A. Beyer, A. Datti, K. Woltjen, A. Nagy, and J. L. Wrana (2010, Jul). Functional genomics reveals a bmp-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell Stem Cell 7*(1), 64–77.

Savageau, M. A. (2001, Mar). Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos 11*(1), 142–159.

Schnerch, A., C. Cerdan, and M. Bhatia (2010, Mar). Distinguishing between mouse and human pluripotent stem cell regulation: the best laid plans of mice and men. *Stem Cells 28*(3), 419–430.

Semenza, G. L. (2000, Aug). Hif-1 and human disease: one highly involved factor. *Genes Dev 14*(16), 1983–1991.

Semenza, G. L. (2003, Oct). Targeting hif-1 for cancer therapy. *Nat Rev Cancer 3*(10), 721–732.

Sengupta, N. and E. Seto (2004, Sep). Regulation of histone deacetylase activities. *J Cell Biochem 93*(1), 57–67.

Seyed, M. and J. X. Dimario (2007, Oct). Sp1 is required for transcriptional activation of the fibroblast growth factor receptor 1 gene in neonatal cardiomyocytes. *Gene 400*(1-2), 150–157.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (2003, Nov). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res 13*(11), 2498–2504.

Shao, Z. and P. D. Robbins (1995, Jan). Differential regulation of e2f and sp1-mediated transcription by g1 cyclins. *Oncogene 10*(2), 221–228.

Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon (2002, May). Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet 31*(1), 64–68.

Shie, J. L., Z. Y. Chen, M. Fu, R. G. Pestell, and C. C. Tseng (2000, Aug). Gut-enriched krüppel-like factor represses cyclin d1 promoter activity through sp1 motif. *Nucleic Acids Res 28*(15), 2969–2976.

Shimizu, Y., T. Takeuchi, S. Mita, T. Notsu, K. Mizuguchi, and S. Kyo (2010, Nov). Krüppel-like factor 4 mediates anti-proliferative effects of progesterone with g_0/g_1 arrest in human endometrial epithelial cells. *J Endocrinol Invest 33*(10), 745–750.

Shmulevich, I. (2002, February). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics 18*(2), 261–274.

Sierra, J., T. Yoshida, C. A. Joazeiro, and K. A. Jones (2006, Mar). The apc tumor suppressor counteracts beta-catenin activation and h3k4 methylation at wnt target genes. *Genes Dev 20*(5), 586–600.

Silva, J., J. Nichols, T. W. Theunissen, G. Guo, A. L. van Oosten, O. Barrandon, J. Wray, S. Yamanaka, I. Chambers, and A. Smith (2009, Aug). Nanog is the gateway to the pluripotent ground state. *Cell 138*(4), 722–737.

Simonsson, S. and J. Gurdon (2004, Oct). Dna demethylation is necessary for the epigenetic reprogramming of somatic cell nuclei. *Nat Cell Biol 6*(10), 984–990.

Singec, I., R. Jandial, A. Crain, G. Nikkhah, and E. Y. Snyder (2007). The leading edge of stem cell therapeutics. *Annu Rev Med 58*, 313–328.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics 19*, 279–281.

Som, A., C. Harder, B. Greber, M. Siatkowski, Y. Paudel, G. Warsow, C. Cap, H. Schöler, and G. Fuellen (2010). The plurinetwork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One 5*(12), e15165.

Student (1908). The probable error of a mean. *Biometrika 6 (1)*, 1–25.

Sun, H. and R. Baserga (2008, Jun). The role of insulin receptor substrate-1 in transformation by v-src. *J Cell Physiol 215*(3), 725–732.

Swarbrick, A., M. C. Akerfeldt, C. S. L. Lee, C. M. Sergio, C. E. Caldon, L.-J. K. Hunter, R. L. Sutherland, and E. A. Musgrove (2005, Jan). Regulation of cyclin expression and cell cycle progression in breast epithelial cells by the helix-loop-helix protein id1. *Oncogene 24*(3), 381–389.

Takahashi, K., K. Tanabe, M. Ohnuki, and M. Narita (2007, November). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell 131*(5), 861–872.

Takahashi, K. and S. Yamanaka (2006, Aug). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell 126*(4), 663–676.

Tapias, A., C. J. Ciudad, I. B. Roninson, and V. Noé (2008, Sep). Regulation of sp1 by cell cycle related proteins. *Cell Cycle 7*(18), 2856–2867.

ten Freyhaus, H., M. Dagnell, M. Leuchs, M. Vantler, E. M. Berghausen, E. Caglayan, N. Weissmann, B. K. Dahal, R. T. Schermuly, A. Ostman, K. Kappert, and S. Rosenkranz (2011, Apr). Hypoxia enhances platelet-derived growth factor signaling in the pulmonary vasculature by down-regulation of protein tyrosine phosphatases. *Am J Respir Crit Care Med 183*(8), 1092–1102.

Terfve, C., T. Cokelaer, D. Henriques, A. MacNamara, E. Goncalves, M. K. Morris, M. van Iersel, D. A. Lauffenburger, and J. Saez-Rodriguez (2012). Cellnoptr: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol 6*, 133.

Thomson, J. P., P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. W. Kerr, A. Deaton, R. Andrews, K. D. James, D. J. Turner, R. Illingworth, and A. Bird (2010, April). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature 464*(7291), 1082–6.

Tokuriki, A., T. Iyoda, K. Inaba, K. Ikuta, S. Fujimoto, M. Kumakiri, and Y. Yokota (2009, Sep). Dual role for id2 in chemical carcinogen-induced skin tumorigenesis. *Carcinogenesis 30*(9), 1645–1650.

Torres, L., J. Sandoval, E. Penella, R. Zaragozá, C. García, J. L. Rodríguez, J. R. Viña, and E. R. García-Trevijano (2009, Jun). In vivo gsh depletion induces c-myc expression by modulation of chromatin protein complexes. *Free Radic Biol Med 46*(11), 1534–1542.

Torres-Padilla, M.-E., D.-E. Parfitt, T. Kouzarides, and M. Zernicka-Goetz (2007, Jan). Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature 445*(7124), 214–218.

Tsuji-Takayama, K., T. Inoue, Y. Ijiri, T. Otani, R. Motoda, S. Nakamura, and K. Orita (2004, Oct). Demethylating agent, 5-azacytidine, reverses differentiation of embryonic stem cells. *Biochem Biophys Res Commun 323*(1), 86–90.

Turkson, J. and R. Jove (2000, Dec). Stat proteins: novel molecular targets for cancer drug discovery. *Oncogene 19*(56), 6613–6626.

Twardziok, S., H. Siebert, and A. Heyl (2010). Stochasticity in reactions : a probabilistic Boolean modeling approach. *Evolution*, 76–85.

Valcourt, U., M. Kowanetz, H. Niimi, C.-H. Heldin, and A. Moustakas (2005, Apr). Tgf-beta and the smad signaling pathway support transcriptomic reprogramming during epithelial-mesenchymal cell transition. *Mol Biol Cell 16*(4), 1987–2002.

Vierbuchen, T., A. Ostermeier, Z. P. Pang, Y. Kokubu, T. C. Südhof, and M. Wernig (2010, February). Direct conversion of fibroblasts to functional neurons by defined factors. *Nature 463*(7284), 1035–1041.

Wada, T., S. Shimba, and M. Tezuka (2006, Jan). Transcriptional regulation of the hypoxia inducible factor-2alpha (hif-2alpha) gene during adipose differentiation in 3t3-l1 cells. *Biol Pharm Bull 29*(1), 49–54.

Waddington, C. (1942). The epigenotype. *Endeavour 1*, 18–20.

Waddington, C. H. (1953). *The epigenetic of birds.* Cambridge University Press.

Waltermann, C., M. Flöttmann, and E. Klipp (2010, Jul). G1 and g2 arrests in response to osmotic shock are robust properties of the budding yeast cell cycle. *Genome Inform 24*(1), 204–217.

Wang, Y. and J. Adjaye (2010, December). A Cyclic AMP Analog, 8-Br-cAMP, Enhances the Induction of Pluripotency in Human Fibroblast Cells. *Stem cell reviews*.

Watson, J. D. (1965). *Molecular biology of the gene.* New York: W. A. Benjamin.

Wei, D., M. Kanai, S. Huang, and K. Xie (2006, Jan). Emerging role of klf4 in human gastrointestinal cancer. *Carcinogenesis 27*(1), 23–31.

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika 34 (1–2)*, 28–35.

Wen, W., J. Ding, W. Sun, K. Wu, B. Ning, W. Gong, G. He, S. Huang, X. Ding, P. Yin, L. Chen, Q. Liu, W. Xie, and H. Wang (2010, Mar). Suppression of cyclin d1 by hypoxia-inducible factor-1 via direct mechanism inhibits the proliferation and 5-fluorouracil-induced apoptosis of a549 cells. *Cancer Res 70*(5), 2010–2019.

Wernig, M., A. Meissner, R. Foreman, T. Brambrink, M. Ku, K. Hochedlinger, B. E. Bernstein, and R. Jaenisch (2007, Jul). In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *Nature 448*(7151), 318–324.

Whitley, D. (1989). The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best. In *Proceedings of the Third International Conference on Genetic Algorithms*, pp. 116–121. Morgan Kaufmann.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 80–83.

Willoughby, J. A., S. N. Sundar, M. Cheung, A. S. Tin, J. Modiano, and G. L. Firestone (2009, Jan). Artemisinin blocks prostate cancer growth and cell cycle progression by disrupting sp1 interactions with the cyclin-dependent kinase-4 (cdk4) promoter and inhibiting cdk4 gene expression. *J Biol Chem 284*(4), 2203–2213.

Wolfraim, L. A., T. M. Walz, Z. James, T. Fernandez, and J. J. Letterio (2004, Sep). p21cip1 and p27kip1 act in synergy to alter the sensitivity of naive t cells to tgf-beta-mediated g1 arrest through modulation of il-2 responsiveness. *J Immunol 173*(5), 3093–3102.

Won, J., J. Yim, and T. K. Kim (2002, Oct). Sp1 and sp3 recruit histone deacetylase to repress transcription of human telomerase reverse transcriptase (htert) promoter in normal human somatic cells. *J Biol Chem 277*(41), 38230–38238.

Wood, A. and A. Shilatifard (2004). Posttranslational modifications of histones by methylation. *Adv Protein Chem 67*, 201–222.

Wu, A., J. Chen, and R. Baserga (2008, Jan). Nuclear insulin receptor substrate-1 activates promoters of cell cycle progression genes. *Oncogene 27*(3), 397–403.

Wu, Q., A. W. Bruce, A. Jedrusik, P. D. Ellis, R. M. Andrews, C. F. Langford, D. M. Glover, and M. Zernicka-Goetz (2009, Nov). Carm1 is required in embryonic stem cells to maintain pluripotency and resist differentiation. *Stem Cells 27*(11), 2637–2645.

Wu, Z., E. D. Rosen, R. Brun, S. Hauser, G. Adelmant, a. E. Troy, C. McKeon, G. J. Darlington, and B. M. Spiegelman (1999, February). Cross-regulation of C/EBP alpha and PPAR gamma controls the transcriptional pathway of adipogenesis and insulin sensitivity. *Molecular cell 3*(2), 151–8.

xin Xie, T., F.-J. Huang, K. D. Aldape, S.-H. Kang, M. Liu, J. E. Gershenwald, K. Xie, R. Sawaya, and S. Huang (2006, Mar). Activation of stat3 in human melanoma promotes brain metastasis. *Cancer Res 66*(6), 3188–3196.

Xu, Q., J. Briggs, S. Park, G. Niu, M. Kortylewski, S. Zhang, T. Gritsko, J. Turkson, H. Kay, G. L. Semenza, J. Q. Cheng, R. Jove, and H. Yu (2005, Aug). Targeting stat3 blocks both hif-1 and vegf expression induced by multiple oncogenic growth signaling pathways. *Oncogene 24*(36), 5552–5560.

Yabut, O. and H. S. Bernstein (2011, May). The promise of human embryonic stem cells in aging-associated diseases. *Aging (Albany NY) 3*(5), 494–508.

Yancopoulos, G. D., P. D. Nisen, A. Tesfaye, N. E. Kohl, M. P. Goldfarb, and F. W. Alt (1985, Aug). N-myc can cooperate with ras to transform normal cells in culture. *Proc Natl Acad Sci U S A 82*(16), 5455–5459.

Yoshida, Y., K. Takahashi, K. Okita, T. Ichisaka, and S. Yamanaka (2009, Sep). Hypoxia enhances the generation of induced pluripotent stem cells. *Cell Stem Cell 5*(3), 237–241.

Yu, J., M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson (2007, Dec). Induced pluripotent stem cell lines derived from human somatic cells. *Science 318*(5858), 1917–1920.

Yu, Q., M. A. Ciemerych, and P. Sicinski (2005, Oct). Ras and myc can drive oncogenic cell proliferation through individual d-cyclins. *Oncogene 24*(47), 7114–7119.

Zaniolo, K., S. Desnoyers, S. Leclerc, and S. L. Guérin (2007). Regulation of poly(adp-ribose) polymerase-1 (parp-1) gene expression through the post-translational modification of sp1: a nuclear target protein of parp-1. *BMC Mol Biol 8*, 96.

Zhang, T., L. B. Nanney, M. O. Peeler, C. S. Williams, L. Lamps, K. J. Heppner, R. N. DuBois, and R. D. Beauchamp (1997, May). Decreased transforming growth factor beta type ii receptor expression in intestinal adenomas from min/+ mice is associated with increased cyclin d1 and cyclin-dependent kinase 4 expression. *Cancer Res 57*(9), 1638–1643.

Zhou, H., S. Wu, J. Y. Joo, S. Zhu, D. W. Han, T. Lin, S. Trauger, G. Bien, S. Yao, Y. Zhu, G. Siuzdak, H. R. Schöler, L. Duan, and S. Ding (2009, May). Generation of induced pluripotent stem cells using recombinant proteins. *Cell stem cell 4*(5), 381–4.

Zhou, Q., H. Chipperfield, and D. Melton (2007, October). A gene regulatory network in mouse embryonic stem cells. *Proceedings of the 104*(42), 16438–16443.

Zhu, X., M. Gerstein, and M. Snyder (2007, May). Getting connected: analysis and principles of biological networks. *Genes Dev 21*(9), 1010–1024.

Zode, G. S., A. F. Clark, and R. J. Wordinger (2009, May). Bone morphogenetic protein 4 inhibits tgf-beta2 stimulation of extracellular matrix proteins in optic nerve head cells: role of gremlin in ecm modulation. *Glia 57*(7), 755–766.

A Appendix

A.1 Microarray Data of Early Reprogramming

Table A.1: Microarray Data of Early Reprogramming FIB:

Shown are the raw data for the different combinations of transduced factors. The first column, without factors, corresponds to the starting point of reprogramming, i.e. fibroblasts in vitro. The other 6 columns show fibroblasts 96 hours *post infectionem* with OCT4, SOX2, KLF4, c-MYC, a combination of the first 3 and all 4 together. All p-values of the measurement satisfy the condition: $p < 0.01$ which is why they are not explicitly shown

Time point	0h FIB	0h GFP	96h	96h	96h	96h	96h	96h
Factors	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC	OCT4 SOX2 KLF4 c-MYC
Treatment	0 0 0 0	0 0 0 0	1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1	1 1 1 0	1 1 1 1
CARM1	5844.701	5687.432	4611.927	4309.676	8818.662	14022.13	7402.918	8499.995
CCND1	22106.14	32344.83	35493.21	32228.53	35338.15	43786.87	40420.59	31561.28
CDK4	3096.276	3245.51	3482.254	2934.213	2921.741	5216.836	2354.626	2996.931
CDK6	1199.035	1239.567	1055.11	1717.921	1163.554	2454.004	862.4547	1690.709
EPAS1	9730.422	6808.151	4479.68	4034.905	3798.041	2810.543	1885.33	974.4784
FGF2	264.4088	444.8094	447.291	490.5566	349.1839	638.3935	201.9951	170.5143
FGFR1	87.63675	114.7827	62.24104	143.649	66.84285	65.70895	70.35942	44.3879
GREM1	1383.041	1288.018	1093.842	1613.278	835.8109	789.4926	670.944	320.2468
GSK3B	861.8276	1097.41	997.6971	691.8084	982.1716	1905.022	1127.087	1354.025
HIF1A	793.8848	1067.593	951.071	1126.612	600.2157	973.3394	502.2162	430.75
ID2	503.179	505.4869	470.4292	1002.207	1060.264	255.3603	504.2882	298.885
ID3	2408.357	2983.359	2080.906	5523.448	4721.275	4512.896	2838.427	2104.678

IRS1	2290.961	2608.995	2046.07	2151.36	1793.798	1364.685	1796.441	810.9319
KLF4	229.6381	204.5731	357.8221	218.7386	113.1707	372.6115	124.612	84.56172
MYC	184.2839	332.7749	346.778	237.9129	199.517	305.2697	184.0885	91.01981
PARP1	3101.696	2736.106	2707.943	2454.829	1966.587	6197.414	3072.92	5349.384
PTPN11	3131.587	3824.121	3688.003	3442.98	3173.27	4964.346	2904.084	2698.67
PTPRU	113.1923	133.2222	101.385	79.5343	98.44271	167.8625	74.65521	75.06278
SMAD3	4188.795	3701.312	3281.479	3230.154	3737.986	3147.293	2248.442	1885.844
SP1	529.7087	414.2288	355.364	314.3633	399.9443	579.725	320.2922	313.8804
STAT3	2523.536	2157.935	2882.218	2708.301	2050.855	1410.546	1695.913	2504.139
TBX3	549.9428	401.9285	303.7661	372.5693	309.4635	452.9981	261.6427	252.4483
TGFBR2	2306.327	2056.21	1604.629	3247.643	1588.182	1184.795	1745.679	1319.628
TLE1	499.6671	415.8826	309.5674	478.0128	304.4097	276.9481	319.2404	254.4583
POU5F1	0	0	3544.585	0	33.15971	0	794.3524	1361.961

A.2 Normalization Procedure of CellNetOptimizer

The exact normalization of the continuous data for the optimization of a Boolean and thus binary model is described in detail in Saez-Rodriguez et al. (2009). In short, the data will be normalized between 0 and 1 but not discretized. First, a fold change relative to a control (*FIB* or *GFP* in our case) is calculated and further modified by a Hill function and a penalty for low signals that are close to the background measurement.

The normalization procedure works as follows:

1. A dynamic range for the measurement of the data is defined by the parameters *detection* and *saturation*. They define the lower bound of sensitivity of the equipment and the upper bound respectively. In our case, these parameters are set to 0 and *infinity* respectively. Whenever a value is outside this dynamic range, it is set to NA.
2. When using modes *time* or *ctrl*, values are transformed into fold changes with respect to time point 0 or the control at the same time and condition respectively. The *raw* mode does not compute fold changes.
3. The fold changes (or the values) are transformed with a Hill function:

$$\frac{x^{HillCoeff}}{EC50Data^{HillfCoeff} + x^{HillCoeff}} \tag{A.1}$$

   where x is the respective data point, *HillCoeff* is the Hill coefficient used for the normalization and *EC50Data* is the normalization parameter corresponding to half-maximal saturation in Hill kinetics.
4. Noisiness of the data is penalized by computing the data value divided by the maximum value of all conditions and times for the species in question
5. The noise penalty is transformed by a saturation function
6. The noise penalty and the result from the Hill function are multiplied
7. If the fold change is negative and bigger than *ChangeTh*, the resulting product is multiplied by -1, if the fold change is smaller than *ChangeTh* (either positive or negative), it is set to 0

A.3 Edge Probabilities of Optimized Models

Table A.2: Edge Probabilities of Optimized Whole Manually Reduced Pluripotency Network

Interaction	**Probability**
SP1 (1) IRS1	1.000
SP1 (1) HIF1A	1.000
SP1 (1) FGFR1	1.000
SP1 (1) EPAS1	1.000
HIF1A (1) FGF2	1.000
POU5F1ext (1) and94	0.995
IRS1 (1) and94	0.995
and94 (1) STAT3	0.995
KLF4ext (-1) SP1	0.602
SP1 (1) MYC	0.568
STAT3 (1) and50	0.519
SP1 (1) and50	0.519
and50 (1) KLF4	0.519
STAT3 (1) KLF4	0.49
STAT3 (1) POU5F1	0.424
MYCext (1) and7	0.405
IRS1 (1) and7	0.405
and7 (1) CCND1	0.405
SP1 (1) and14	0.389
MYCext (1) and14	0.389
and14 (1) CCND1	0.389
MYCext (1) and10	0.319
KLF4ext (-1) and10	0.319
and10 (1) CCND1	0.319
SP1 (1) and64	0.314
SMAD3 (-1) and64	0.314
and64 (1) MYC	0.314
SP1 (1) and55	0.294
ID2 (-1) and55	0.294
and55 (1) MYC	0.294
STAT3 (1) and100	0.189
EPAS1 (1) and100	0.189
and100 (1) POU5F1	0.189
STAT3 (1) and109	0.164
POU5F1ext (1) and109	0.164
and109 (1) POU5F1	0.164
KLF4 (1) POU5F1	0.162
KLF4 (1) and97	0.151
EPAS1 (1) and97	0.151
and97 (1) POU5F1	0.151
STAT3 (1) and107	0.116
KLF4 (1) and107	0.116

and107 (1) POU5F1	0.116
POU5F1ext (1) and98	0.108
EPAS1 (1) and98	0.108
and98 (1) POU5F1	0.108

Table A.3: Edge Probabilities of Optimized Minimal Network for Early Reprogramming

Interaction	**Probability**
SP1 (1) EPAS1	1.000
SP1 (1) HIF1A	1.000
SP1 (1) MYC	1.000
SP1 (1) IRS1	1.000
SP1 (1) FGFR1	1.000
HIF1A (1) FGF2	1.000
POU5F1ext (1) and49	1.000
IRS1 (1) and49	1.000
and49 (1) STAT3	1.000
KLF4ext (-1) SP1	0.701
STAT3 (1) KLF4	0.690
SP1 (1) and14	0.594
MYCext (1) and14	0.594
and14 (1) CCND1	0.594
STAT3 (1) and40	0.407
SP1 (1) and40	0.407
and40 (1) KLF4	0.407
KLF4 (1) POU5F1	0.393
STAT3 (1) POU5F1	0.303
MYCext (1) and10	0.290
KLF4ext (-1) and10	0.290
and10 (1) CCND1	0.290
MYCext (1) and7	0.258
IRS1 (1) and7	0.258
and7 (1) CCND1	0.258
STAT3 (1) and63	0.226
POU5F1ext (1) and63	0.226
and63 (1) POU5F1	0.226
POU5F1ext (1) and52	0.154
EPAS1 (1) and52	0.154
and52 (1) POU5F1	0.154
KLF4 (1) and51	0.148
EPAS1 (1) and51	0.148
and51 (1) POU5F1	0.148

STAT3 (1) and54	0.132
EPAS1 (1) and54	0.132
and54 (1) POU5F1	0.132
POU5F1ext (1) and59	0.127
KLF4 (1) and59	0.127
and59 (1) POU5F1	0.127

Table A.4: Edge Probabilities of Optimized Minimalistic Pluripotency Network With SP1

Interaction	**Probability**
SP1 (1) HIF1A	1.000
POU5F1ext (1) and52	1.000
KLF4 (1) and52	1.000
and52 (1) POU5F1	1.000
SP1 (1) MYC	0.964
SP1 (1) KLF4	0.526
KLF4ext (-1) SP1	0.508
SP1 (1) and11	0.455
MYCext (1) and11	0.455
and11 (1) CCND1	0.455
MYCext (1) and12	0.341
MYC (1) and12	0.341
and12 (1) CCND1	0.341
MYCext (1) and6	0.326
KLF4ext -1 and6	0.326
and6 (1) CCND1	0.326